N°.D'ORDRE :15/2010-M/EL

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université des Sciences et de la Technologie Houari Boumediene

Faculté d'Electronique et Informatique

présenté pour l'obtention du diplôme de MAGISTER en

ELECTRONIQUE

Spécialité: Communication parlée

Par : **Mohammed SAIDI**

Codage Sinusoïdal de la Parole pour des Transmissions de la Voix sur IP (VoIP)

Soutenu publiquement le : 25/05/2010, devant le jury composé de:

Mr. B.HADDAD	**Professeur à l'USTHB**	**Présedent de jury**
Mr. B.BOUDRAA	**Maître de Conférences A à l'USTHB**	**Directeur de thèse**
Mr. A.AMROUCHE	**Maître de Conférences A à l'USTHB**	**Examinateur**
Mr. M.BOUZID	**Maître de Conférences A à l'USTHB**	**Examinateur**
Mr. H.TEFFAHI	**Maître de Conférences A à l'USTHB**	**Examinateur**

Dédicaces

A mes chers parents, à toute ma famille et tous mes amis.

Remerciements

Tout d'abord, je remercie Dieu le Tout puissant de m'avoir gardé en santé et donner la persévérance et le courage durant la période d'accomplissement de ce travail.

Ce travail a été réalisé au sein du Laboratoire de Communication Parlée et Traitement du Signal (LCPTS) de la faculté d'Electronique et d'Informatique de l'USTHB.

J'adresse tout particulièrement ma reconnaissance à mon directeur de thèse : Monsieur Bachir BOUDRAA, Maître de Conférences à l'U.S.T.H.B, pour les nombreux conseils et suggestions qu'il n'a cessé de me donner et qui m'ont été très précieux tout au cours de la réalisation de ce mémoire. Je tiens à remercier pour la confiance et la sympathie qui m'a témoignées, aussi je lui exprime ma profonde gratitude pour sa disponibilité constante, son aide et sa patience avec moi.

Je remercie les membres de jury qui ont accepté d'examiner ce travail et d'y apporter leur témoin :

Monsieur Boualem HADDAD, Professeur à l'université de l'U.S.T.H.B, qui me fait le grand honneur d'accepter la présidence du jury.

Monsieur Abderrahmane AMROUCHE, Maître de Conférences à l'université de l'U.S.T.H.B, pour l'honneur qu'elle me fait en acceptant de participer à ce jury.

Monsieur Merouane BOUZID, Maître de Conférences à l'université de l'U.S.T.H.B, pour l'honneur qu'elle me fait en acceptant de participer à ce jury.

Monsieur Hocine TEFFAHI, Maître de Conférences à l'université de l'U.S.T.H.B, pour l'honneur qu'elle me fait en acceptant de participer à ce jury.

Je tiens également à faire part de toute ma gratitude et ma sympathie à tous les membres de l'équipe (enseignants et étudiants) de notre laboratoire de communication parlée.

Mes vifs remerciements à mes chers parents que j'aime beaucoup, qui ont toujours été présent lorsque j'en ai eu besoin, à toute ma famille et à tous mes amis.

Enfin, j'exprime toute ma gratitude à ceux qui, de près ou de loin, chacun à sa manière, ont contribué à l'élaboration de mémoire.

Résumé

La transmission de la voix sur Internet (VoIP) se fait par émission et réception de paquets. Au récepteur, certains paquets manquent à cause des délais, de la congestion ou des erreurs de transfert. Cette perte de paquets dégrade la qualité de la voix au niveau du récepteur. Des algorithmes de dissimulation des pertes (concealing) sont utilisés au niveau de l'émetteur ou au niveau du récepteur afin de combler ces pertes.

Dans ce travail, nous avons mis au point et implémenté un codeur harmonique, véhiculant la voix sur IP. Ce codeur est conçu en utilisant la prédiction linéaire à excitation mixte (MELP).

Etant donné que la transmission de la voix est effectuée en temps réel, le récepteur ne peut pas requérir à la retransmission des paquets perdus à cause des délais de transfert trop importants. Nous avons alors utilisé la technique de codage par description multiple (MDC), qui est basée sur la redondance des informations. Ce sont ces redondances que l'on appelle descriptions. Ces dernières sont transportées dans chaque paquet à partir duquel les trames perdues seront reconstituées. Ainsi, nous avons codé le signal sur deux descriptions : la première permet de coder la trame courante à 2.4 kbps. Elle sert à la reconstruction du signal avec une bonne qualité. La seconde description utilise un autre codeur MELP, mais fonctionnant à un débit de 1.2 kbps. Ce dernier code trois trames successives dans le même paquet, à savoir les trois trames qui succèdent la trame courante. Cette description contribue à reconstruire la parole lorsqu'un, deux, voire trois paquets seront perdus. Nous pouvons ainsi récupérer jusqu'à trois trames lorsqu'on perd trois paquets successifs. L'objectif est de garder l'intelligibilité et une certaine qualité de la parole lorsque le taux de perte de paquets augmente.

Pour nos simulations, les pertes de paquets se font de façon aléatoire. A la réception, et dans le cas sans perte, on forme la voix à partir de la première description (codeur MELP à 2.4 kbps). En cas de perte, la seconde description (MELP à 1.2 kbps) reçue précédent, et préalablement mémorisée au niveau du récepteur, servira à la reconstitution.

Pour valider notre méthode, nous avons utilisé un matériau linguistique formé de corpus multilingue. Le premier est composé de phrases arabes phonétiquement équilibrées conçues au niveau de notre laboratoire. Pour les langues française et anglaise, nous avons utilisé les phrases célèbres phonétiquement équilibrées : « la bise et le soleil » et « the sun and wind ». L'évaluation de la qualité de la parole utilise la méthode PESQ, normalisée ITU-T, conjointement à des tests d'écoute. Les résultats obtenus montrent que la méthode proposée améliore considérablement la qualité de la voix reconstruite lorsque celle-ci est véhiculée sur IP.

Table des matières

Liste des Figures

Liste des Tableaux

Liste des abréviations

ACELP	Algebric CELP (prédiction linéaire excite par les séquences codées à structure algébrique
AMR	Adaptive Multi-Rate
ARMA	Auto Régressif à Moyenne Ajustée
ATM	Asynchrone Transforme Mode
CELP	Code Excited Linear Prediction (prédiction linéaire avec excitation par code)
CAN	Convertisseur analogique numérique
CNA	Convertisseur numérique analogique
DoD	Department of Defense
DSL	Digital Subscriber Line
DFT	Discrete Fourier Transform(TFD : Transformation de Fourier Discrete)
DSP	Digital Signal Processing
EQM	Erreur Quadratique Moyenne
FEC	Forward Error Correction
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
GSM	Global System Mobile
IETF	Internet Engineering Task Force
IP	Internet protocole
ITU	International Telecommunication Union (Union International des Télécom)
iLBC	Internet Low Bit Rate
LAN	Local Area Network
LD-CELP	Low-delay CELP (CELP à délai réduit)
LP	Linear Prediction (Prédiction Linéaire)
LPC	Linear Predictive Coding (Codage de prédiction linéaire)
LSF	Line Spectral Frequencies (Fréquences de raies spectrales)
LTP	Long-Term Prediction (Prédiction à Long Terme)
MDC	Multiple Description Coding
MELP	Excitation Linear Prediction (Excitation Mixte Prediction Linéaire)
MDSQ	Multiple Description Scalar Quantization
MDVQ	Multiple Description Vector Quantization
MIC	Modulation par Impulsion et Codage
MOS	Mean Opinion Score
MSVQ	Multistage Vector Quantization
PAPE	Phrases Arabes Phonétiquement Equilibrées
PCM	Pulse Code Modulation.
PESQ	Perceptual Evaluation of Speech Quality
PLC	Packet Loss Concealment
RMS	Root Mean Square
RSB	Rapport Signal sur Bruit
RTC	Réseau Téléphonique Commuté

RTCP	Real-Time Transport Control Protocol
RTP	Real-Time Transport Protocol
SD	Spectral Distortion
TCM	Trellis Coded Modulation
TCP	Transport Control Protocol
TCQ	Trellis Coded Quantization
ToIP	Telephony over IP
QoS	Quality of Service.
UDP	User Datagram Protocol
UEP	Unequal Error Protection (Protection inégale contre les erreurs)
VAD	Voice Activity Detection
VoIP	Voice over IP
VSELP	Vector-Sum Excited Linear Prediction.
V/UV	Voiced /Unvoiced
QS	Quantification scalaire
QV	Quantification vectorielle

Introduction générale

La parole représente encore le plus grand trafic véhiculé par les réseaux de télécommunications. Jusqu'à une époque récente, pour transmettre un signal de parole, des chaînes de communication analogiques ont été utilisées presque exclusivement. A cause des perturbations et des bruits apparaissant inévitablement dans le canal de transmission, le signal reconstruit au récepteur ne pouvait être une réplique exacte du signal émis.

La numérisation actuelle du réseau de télécommunication a permis d'améliorer considérablement la qualité des signaux véhiculés. Le codage numérique de la parole est aujourd'hui présent sur la plupart des chaînes de communication.

Jusqu'au début des années 80, l'utilisation du traitement du signal à grande échelle sur les réseaux s'est limitée à la quantification logarithmique du codage MIC à 64 kbps pour la parole de nature téléphonique. Pour économiser les ressources des canaux de communication, le signal numérisé devait alors être compressé pour réduire le flux de données nécessaires à une reconstruction de bonne qualité du signal vocal d'origine.

Avec le développement de calculateurs de plus en plus performants, le traitement numérique du signal a remplacé de nos jours le traitement analogique et a rendu possible la mise en œuvre d'algorithmes de plus en plus précieux pour la compression du signal de parole. Avant tout algorithme, il est nécessaire de numériser les signaux. Le débit binaire du signal numérisé est alors égal au produit de la fréquence d'échantillonnage par le nombre d'éléments binaires nécessaires à la représentation de toutes les valeurs discrètes du signal. Pour réduire ce débit, divers algorithmes ont permis de diminuer les redondances inutiles du signal, en vue de sa transmission à travers un canal numérique. Ce dernier peut être un canal cellulaire numérique, un canal satellite Internet.

Dans le monde moderne des télécommunications, la voix sur le réseau IP (VoIP : Voice over Internent Protocol) a pris une importance considérable ces dernières années. La tendance récente est de remplacer les réseaux de commutation de circuits, tels que le RTC, par une transmission de la voix par des paquets sur les réseaux Internet. Cette dernière exigeant des codeurs de parole de posséder plus de robustesse et de flexibilité pour effectuer des communications vocales de haute qualité. En effet, la transmission de la voix sur réseau IP se fait par envoi de paquets. Au récepteur, la parole transmise subit de nombreuses dégradations dues à de nombreuses raisons. La première d'entre-elles est la distorsion introduite par le codeur de parole utilisé. La seconde est due à la perte de certains paquets à cause des délais, de la congestion (fuite, encombrement,…) ou des erreurs de transfert. Etant donné que la

transmission de la voix est effectuée en temps réel, le récepteur ne peut pas requérir à la retransmission des paquets perdus car les délais de transfert sont jugés trop importants. Des algorithmes de dissimulation des pertes (concealing) sont alors utilisés au niveau de l'émetteur ou au niveau du récepteur afin de combler ces pertes et garder l'intelligibilité de la parole perçue.

L'étude principale, présentée dans ce mémoire, est d'abord la conception et l'implémentation d'un codeur de type harmonique. Il s'agit d'un codeur LPC à excitation mixte (MELP) fonctionnant à un débit de 2.4 kbps. Le choix de ce type de codage est d'assurer d'une part un débit plus faible que le codeur CELP utilisé actuellement pour une telle transmission (8 kbps pour la norme G.729). D'autre part, est d'assurer une robustesse contre les pertes de paquets lors d'une transmission de la voix sur IP. A cet effet, nous proposons d'implanter un algorithme de dissimulation des pertes dans le paquet courant en prenant en considération le contenu des paquets à venir. Nous employons à cet effet, une méthode dite de description multiple, de plus en plus populaire, pour combattre les pertes de paquets et augmenter ainsi la robustesse des systèmes face à ces pertes. Cette multi-description contiendra dans un même paquet deux codeurs MELP à la fois. Le premier fonctionnant à 2.4 kbps servira la bonne transmission de voix. Le second sera utilisé pour recouvrer les éventuelles pertes des paquets.

Organisation du manuscrit

Le mémoire s'articule autour de cinq chapitres.

Le **chapitre I** est consacré à une présentation de l'état de l'art sur le codage de la parole utilisant citant les différents algorithmes à base de prédiction linéaire. Parmi les algorithmes rencontré dans la littérature, nous présentons le MPLPC (codage LPC à excitation multi-impulsionnelle), le CELP (codage LPC à excitation par code) et enfin le codage MELP (codage LPC à excitation mixte), qui le cœur de notre présent travail.

Le **chapitre II** sera consacré à présentation de la voix sur IP.

Le **chapitre III** est consacré à la théorie de la description multiple.

Le **chapitre IV**, présentera notre réalisation effectuée à base de deux codeurs MELP : le premier est le standard MELP fonctionnant à 2.4 kbps, le second est le MELP opérant à 1.2 kbps.

Le **chapitre V**, sera consacré à l'évaluation de ces deux codeurs et à la simulation effectuée utilisant la MDC.

Enfin, nous terminons par une conclusion générale sur le travail accompli ainsi nous donnons les perspectives futures qui peuvent enrichir ce travail.

Chapitre I

Etat de l'art du codage de la parole par prédiction linéaire

I.1.Introduction

Le codage de la parole permet la réduction de débit de transmission du signal et des communications dans des canaux à largeur de bande limitée. La largeur de bande d'une transmission devra être minimisée tout en préservant la qualité du signal vocal reconstruit et en répondant aux autres exigences liées à l'application. Dans le cas de la transmission de la voix sur IP, la réduction du débit limitera le nombre ou la taille des paquets à envoyer sur le réseau.

Les techniques de traitement de signal, utilisées pour le codage de parole, peuvent être basées sur un traitement dans les domaines du temps ou des fréquences. Ces deux types de codeurs, temporel ou fréquentiel, exploitent les redondances du signal vocal pour améliorer leur efficacité. Différentes techniques peuvent être rencontrées. Certains codeurs modélisent le système de production de la parole par extraction de paramètres décrivant le signal vocal. D'autres essaient de réduire au minimum l'erreur entre le signal original et les formes d'onde de la parole reconstruite. Enfin, plusieurs codeurs bas débit emploient une combinaison des deux techniques [1].

I.2. Système de codage

Un système de codage de la parole comprend deux parties : le codeur et le décodeur. Le codeur analyse le signal pour en extraire un nombre réduit de paramètres pertinents qui sont représentés par un nombre restreint de bits pour archivage ou transmission. Le décodeur utilise ces paramètres pour reconstruire un signal de parole synthétique.

La plupart des algorithmes de codage mettent à profit un modèle linéaire simple de production de la parole. Ce modèle sépare la source d'excitation du canal vocal. L'éxcitation peut être quasi périodique pour les sons voisés ou de type bruit pour les sons fricatifs ou plosifs (occlusifs). Le conduit vocal est considéré comme un résonateur acoustique. Sa forme détermine ses fréquences de résonance et l'enveloppe spectrale (formants) du signal de parole.

Le signal de parole est souvent modélisé (modèle « source-filtre ») comme la sortie d'un filtre tout pôle (appelé filtre de synthèse) dont la fonction de transfert représente l'enveloppe spectrale. Ce filtre est excité par une entrée dont les caractéristiques (en particulier la fréquence fondamentale) déterminent la structure fine du spectre [2]. C'est le cas des codeurs à prédiction linéaire.

Le signal de parole n'étant pas stationnaire, les codeurs le découpent généralement en trames quasi-stationnaires de durées comprises entre 5 et 30 ms. Sur chaque trame, le codeur extrait des paramètres représentant l'enveloppe spectrale et caractérise ou modélise l'excitation de manière plus ou moins fine soit par quantification vectorielle, soit à l'aide de paramètres tels que l'énergie, le voisement et la fréquence fondamentale F_0. D'autres paramètres peuvent être calculés pour représenter plus finement l'excitation. Les paramètres les plus souvent utilisés pour l'enveloppe spectrale sont les paires de raies spectrales ou *LSF* (Line Spectral Frequencies) qui sont déduites des coefficients de prédiction linéaire et qui possèdent de bonnes propriétés pour la quantification et l'interpolation.

I.3. Codage de la parole

Afin de coder la parole, plusieurs étapes sont nécessaires. Le signal subit tout d'abord un filtrage anti-repliement, puis un échantillonnage suivi d'une quantification et enfin le codage.
L'échantillonnage est le processus de représentation d'un signal continûment variable par une séquence de valeurs. La quantification consiste à représenter approximativement chaque échantillon dans un ensemble fini de valeurs. Enfin, le codage consiste à assigner un numéro réel à chaque valeur.

I.3.1. Filtrage anti-repliement

Avant l'échantillonnage, un filtre passe-bas de fréquence de coupure égale à la moitié de la fréquence d'échantillonnage est inséré pour éviter l'effet dénommé « repliement » ou « aliasing » postulé par le théorème de Nyquist-Shannon. Ce filtre est appelé filtre« anti-repliement » ou « anti-aliasing ».

I.3.2. L'échantillonnage

L'échantillonnage transforme le signal à temps continu $x(t)$ en un signal à temps discret $x(nTe)$défini aux instants d'échantillonnage, multiples entiers de la période d'échantillonnage *Te*. Celle-ci est elle-même l'inverse de la fréquence d'échantillonnage *Fe*.

En ce qui concerne le signal vocal, le choix de Fe résulte d'un compromis. Son spectre peut s'étendre jusque 12 kHz [3]. Il faut donc en principe choisir une fréquence *Fe* égale à 24 kHz au moins pour satisfaire raisonnablement au théorème de Shannon [4].

Cependant, le coût d'un traitement numérique, filtrage, transmission, ou simplement enregistrement, peut être réduit d'une façon notable si l'on accepte une limitation du spectre par un filtrage préalable. C'est le rôle du filtre de garde, dont la fréquence de coupure *fc* est choisie en fonction de la fréquence d'échantillonnage retenue.

I.3.3. Quantification

Au cours du traitement du signal de parole, toutes les données sont représentées, sur un certain nombre d'éléments binaires, avec une précision finie. Cette opération consiste à représenter le signal analogique, à des instants discrets dans le temps, par une valeur choisie parmi un ensemble fini. Outre la nécessite de la quantification pour numériser les données, elle est aussi un moyen de compression.

I.3.3.1. Quantification scalaire

Considérons un signal à temps discret $x(n)$ prenant ses valeurs dans l'intervalle $[-A, A]$.

Définir un quantificateur scalaire avec une résolution de b bits par échantillon consiste à réaliser trois opérations :

1. Une partition de l'intervalle $[-A, A]$ en $L = 2^b$ intervalles distincts $\{\Theta^1 \dots \Theta^L\}$ de longueur $\{\Delta^1 \dots \Delta^L\}$
2. Une numérotation des éléments de la partition $\{i^1 \dots i^L\}$
3. La sélection d'un représentant par intervalle. L'ensemble de ces représentants forme un dictionnaire (codebook).

La procédure d'encodage (à l'émetteur) consiste à décider à quel élément de la partition appartient $x(n)$ puis à lui associer le numéro $i(n) \in \{1, \dots, L = 2^b\}$ correspondant. C'est le numéro de l'intervalle choisi, qui sera transmis où stocké. La procédure de décodage (au récepteur) consiste à associer au numéro $i(n)$ le représentant correspondant

$\hat{x}(n) = \hat{x}^{\,i}(n)$ choisi parmi l'ensemble des représentants $\{\hat{x}_1 \dots \hat{x}_L\}$. Formellement, on peut observer que la quantification scalaire est une application non bijective de $[-A, A]$ dans un ensemble fini C plus une règle d'affectation.

C'est un processus irréversible entraînant une perte d'information, une erreur de quantification que l'on notera. Il est nécessaire de définir une mesure de distorsion, son expression est donnée dans (1.1). On choisira par la suite la mesure de distorsion la plus petite de l'erreur quadratique, la figure I.1 représente le schéma bloc d'une quantification scalaire.

$$d[x(n), \hat{x}(n)] = [x(n), \hat{x}(n)]^2 \tag{1.1}$$

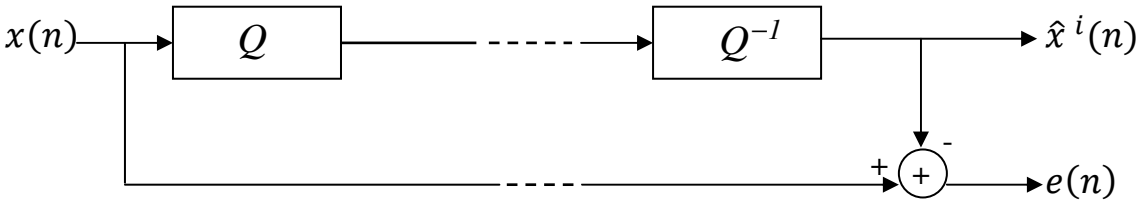


Fig. I.1. Quantification scalaire

L'opération de quantification apportera toujours des dégradations irréversibles par rapport au signal original qui se traduisent par une erreur, dite bruit de quantification $e(n)$ telle que :

$$e(n) = x(n) - \hat{x}^{\,i}(n) \tag{1.2}$$

I.3.3.2. Quantification Vectorielle

La quantification vectorielle ou méthode de compression de données, a pris une place très importante dans le domaine de la communication, que ce soit dans un but de transmission ou d'archivage d'informations. Cette méthode s'applique essentiellement dans les deux domaines que sont l'image et la parole, ces deux formes de signaux contenant des informations redondantes.

La quantification vectorielle est un processus d'approximation d'un signal d'amplitude continue par un signal d'amplitude discrète. Le but de la compression étant d'extraire une information maximale tout en ne créant qu'un minimum de distorsion par rapport au signal original. L'objectif sera de créer une distorsion minimale pour un taux de compression donné.

Cette méthode figure parmi les méthodes qui font l'objet de recherches toujours plus élaborées.

La quantification vectorielle est une généralisation de la quantification scalaire. On appelle quantificateur vectoriel de dimension N et de taille L une application de R^N dans un ensemble fini C contenant L vecteurs de dimension N.

Ces paramètres sont reliés par la relation $L = 2^{bN}$. Autrement dit, le débit binaire b d'un QV utilisant un dictionnaire, comme montré dans la figure I.2, est défini par l'équation suivante :

$$b = \frac{1}{N} log_2 L \tag{1.3}$$

Cette expression du débit permet de faire des études comparatives de quantification opérant sur des vecteurs de dimensions différentes. La figure I.2 illustre le principe de quantification vectorielle [5].

Le quantificateur vectoriel permet de prendre en compte directement la corrélation contenue dans le signal plutôt que de chercher d'abord à décorréler le signal puis à quantifier un signal décorrélé comme le fait le quantificateur scalaire.

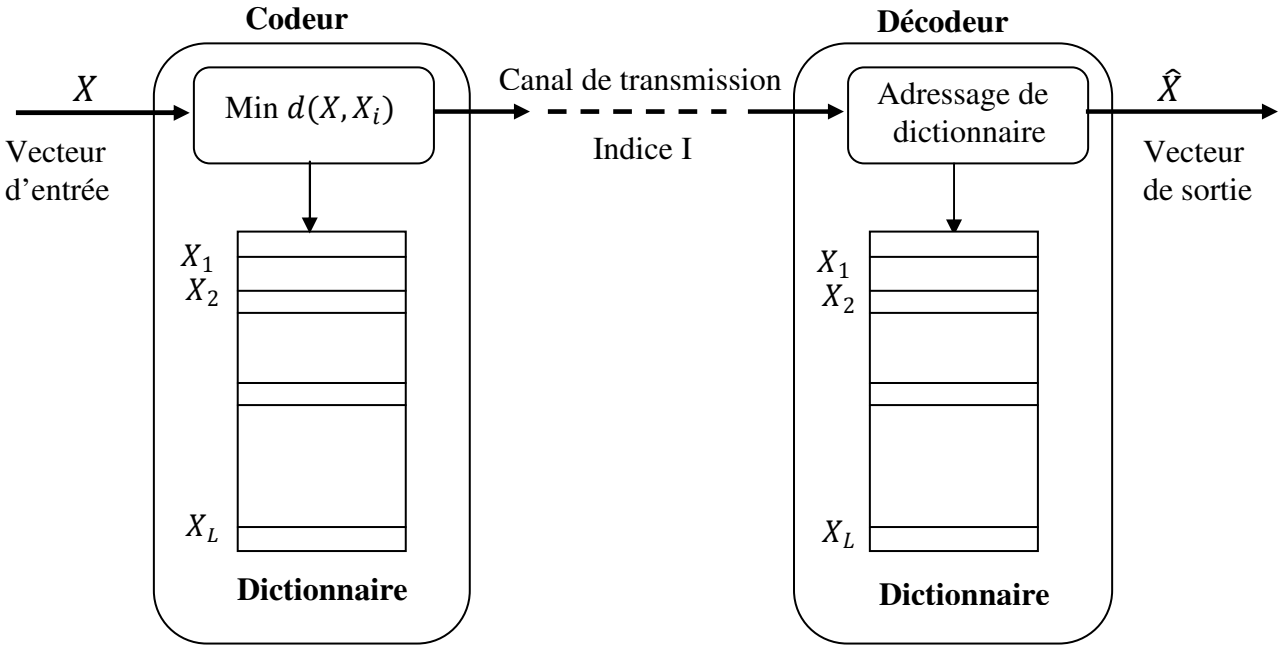


Fig. I.2. Schéma synoptique de la quantification vectorielle

Codeur :

Le rôle du codeur consiste, pour tout vecteur X_i du signal en entrée à rechercher dans le dictionnaire, le vecteur code le plus proche de vecteur source X. C'est uniquement l'adresse du

vecteur code ainsi sélectionnée qui sera transmise ou stockée. C'est à ce niveau donc que s'effectue la compression.

Décodeur :

Il dispose d'une réplique du dictionnaire et consulte celui-ci pour fournir le vecteur code d'indice correspondant à l'adresse reçue. Le décodeur réalise l'opération de décompression.

Dans le codeur, on associe au vecteur X un mot du code X_i du dictionnaire selon le critère du plus proche voisin. Seul l'indice I est transmis au décodeur. On a NC indices correspondant aux NC mots de code du dictionnaire. le même dictionnaire est utilisé au décodage. Le mot de code, représentant du vecteur X, est retrouvé à partir de l'indice I reçu.

I.4. Différentes approches en codage de la parole

Les méthodes de codage de la parole sont nombreuses et sont classées généralement en trois grandes catégories : les codeurs de forme d'onde, les codeurs paramétriques, appelés vocodeurs, et les codeurs hybrides. Le choix d'une méthode va dépendre surtout de l'application visée et des contraintes sur le débit. Les codeurs de forme d'onde sont surtout performants pour des débits élevés (au-delà des 16 kbps). Les vocodeurs quant à eux sont plutôt destinés aux bas débits (2.4 à 8 kbps) et aux très bas débits (au-dessous de 2.4 kbps). Les codeurs hybrides ont des débits nominaux intermédiaires (8 à 16kbps), bien qu'ils puissent aussi être utilisés pour de bas débits. Une autre façon de distinguer les codeurs de parole est la largeur de bande codée. Historiquement, les codeurs de parole sont à bande étroite, c.-à-d. codant le signal sur la bande 300 à 3400 Hz. Le signal d'entrée est alors échantillonné à 8 kHz. Ces dernières années, la tendance est d'augmenter la largeur de bande vers la bande élargie de 50 à 7000 Hz en utilisant une fréquence d'échantillonnage de 16 kHz. La qualité de la synthèse et de la communication s'en trouve nettement améliorée [5].

I.4.1. Codeurs de forme d'onde

Les codeurs de forme d'onde s'efforcent de préserver l'allure temporelle du signal de parole. Le signal reconstruit avec ce type de codeur converge vers le signal original avec l'augmentation du débit de transmission. La qualité du signal synthétisé obtenue est excellente pour un débit relativement élevé. Les premiers codeurs temporels de type PCM (Pulse Code Modulation), qui reposent exclusivement sur le théorème d'échantillonnage de Shannon [6] et une quantification fixe, sont apparus dans les années 60. Etant donnée la distribution d'amplitudes des échantillons de parole, un quantificateur non-uniforme

apporterait une meilleure qualité pour le même débit. Ainsi, l'Union Internationale des Télécommunications (UIT) normalisé le codeur G.711 en 1972, un codeur logarithmique de parole de type PCM pour la transmission téléphonique avec un débit de 64 kbps.

I.4.2. Codeurs paramétriques

Les codeurs paramétriques utilisent le modèle de production et les caractéristiques de la perception humaine pour décrire le signal de parole par un jeu de paramètres. Ce jeu ne permet pas de reconstruire la forme d'onde, mais il permet de synthétiser un signal perceptiblement similaire au signal d'origine. Ainsi, en augmentant le débit, le signal synthétisé ne converge pas vers la forme du signal original et sa qualité est limitée par la précision du modèle. Ces codeurs ont été conçus pour des applications à bas débit et sont principalement prévus pour maintenir l'intelligibilité du signal vocal. La plupart des codeurs paramétriques sont basés sur le codage par prédiction linéaire (LPC). Le plus simple des codeurs LPC est le codeur pour lequel l'excitation est générée par un train d'impulsions espacées de périodes de pitch pour les sons voisés et par un bruit aléatoire pour les sons non-voisés [1]. Dans le modèle de l'excitation mixte, le signal d'excitation est composé d'un train d'impulsions dans les fréquences basses et d'un bruit dans les fréquences hautes. Ce modèle a été élaboré par McCree et Bamwell dans le codeur LPC à excitation mixte (MELP : Mixed Excitation Linear Prediction) où le mélange de la composante harmonique et de la composante de bruit se fait à l'aide de deux filtres RIF (réponse impulsionnelle finie) variables dans le temps [2,6].

I.4.3. Codeurs hybrides

Les codeurs hybrides utilisent les deux méthodes temporelle et paramétrique de façon complémentaire, ce qui permet un codage de parole de bonne qualité à des débits moyens. Ces codeurs sont basés sur des techniques de codage temporel auxquelles des modèles de production de parole sont associés pour améliorer leur efficacité. Cependant, ce type de codage nécessite des coûts de calculs plus importants. Tous les codeurs hybrides s'appuient, eux aussi, sur une analyse LPC pour obtenir les modèles de synthèse de parole. Les deux techniques paramétrique et temporelle modélisent respectivement le conduit vocal et le signal d'erreur résiduel. En 1982, Atal et Remde [1] utilisent le principe d'analyse par synthèse et modélisent le signal d'erreur à partir d'excitations multi-impulsionnelles. Ce n'est qu'en 1985, qu'Atal et Schroeder [7] définissent le codeur CELP (Code Excited Linear Prediction), qui détermine une forme d'onde optimale du signal d'erreur en utilisant l'analyse par synthèse.

I.5. Codeurs de parole à bas débit

Le codage de la parole est une discipline qui a subi une forte évolution depuis 1982, par l'avènement d'un nouveau concept de codage du signal d'excitation. Ce nouveau concept appelé modélisation par analyse-synthèse a vu l'introduction successive du codeur à excitation multi-impulsionnelle (MP-LPC) puis du codeur à excitation par code (CELP) et ensuite le codeur à excitation mixte (MELP).

Le codage prédictif de la parole englobe la plupart des techniques utilisées actuellement pour des débits allant de 16 kbps à 5 kbps même en dessous. Cependant, différentes mises en œuvre du codage prédictif sont possibles. On distingue notamment les méthodes d'analyse et synthèse, de celles d'analyse par synthèse.

I.6. Modèle prédictif de production de la parole

La parole peut être considérée comme étant un signal pseudo-stationnaire, c.-à-d. stationnaire sur de courtes durées allant en général de 5 jusqu' à 30 ms. Sur cette période, il est possible de caractériser le spectre du signal par deux attributs :

- L'enveloppe spectrale.
- La structure fine du spectre.

Le codage linéaire de prédiction se fonde sur un modèle fortement simplifié pour la production de la parole [7].

Un signal voisé peut être modélisé par le passage d'un train d'impulsion $u(n)$ à travers un filtre numérique récursif de type tous pôles. Cette modélisation reste valable dans le cas des sons non voisé, à condition que $u(n)$ soit un bruit blanc. Le modèle final est illustré à la figure I.3. Il est souvent appelé modèle auto-régressif (AR), parce qu'il correspond dans le domaine temporel à une régression linéaire de la forme :

$$\hat{s}(n) = G.u(n) + \sum_{i=1}^{p} -a_i \hat{s}(n-i) \tag{1.4}$$

Où $u(n)$ est le signal d'excitation, ce qui exprime que chaque échantillon est obtenu en ajoutant un terme d'excitation à une prédiction obtenue par combinaison linéaire de P échantillons précédents.

Les coefficients du filtre sont appelés coefficients de prédiction et le modèle AR est souvent appelé modèle de prédiction linéaire.

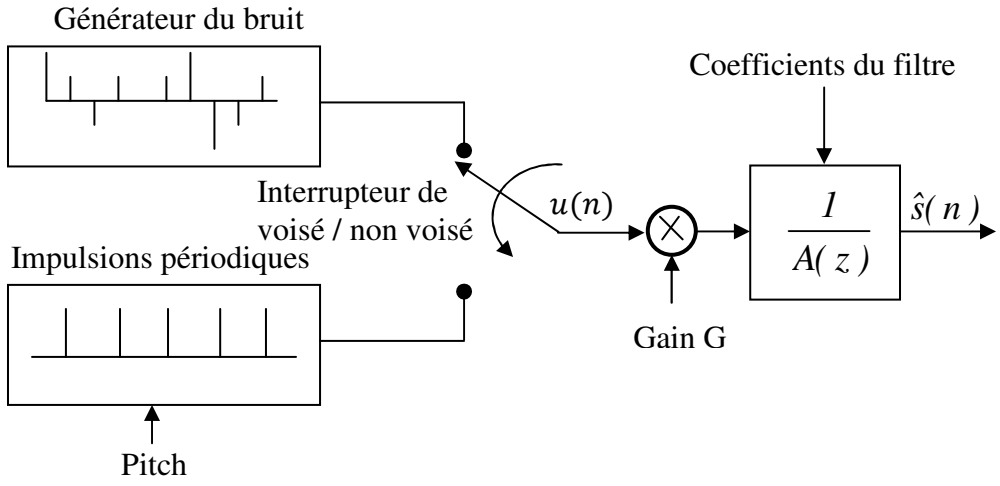


Fig. I.3. Modèle simplifié de production de la parole

Ce modèle comprend :

- Un générateur périodique d'impulsions ;
- Un générateur de bruit blanc ;
- Un interrupteur servant à choisir les sons voisés ou non voisés ;
- Un gain G proportionnel à la valeur efficace du signal ;
- Un filtre tous pôles $H(z) = 1/A(z)$.

Par analogie entre le modèle physique et le modèle mathématique, ont peut donner les relations d'équivalence suivantes [8]:

Conduit vocal	⟺	Filtre LP $1/A(z)$
Flux d'air	⟺	Signal d'excitation ou signal résiduel
Vibration des cordes vocales	⟺	Son voisé
Période de vibration des cordes vocales	⟺	Période du pitch
Volume d'air	⟺	Gain

I.7. Codage basé sur la prédiction linéaire

La prédiction linéaire (LP) fait partie intégrante de la plupart des algorithmes de codage de la parole. Le codeur à prédiction linéaire (LPC) est fondé sur le modèle de la production de la parole, présenté sur la figure I.4. L'excitation est composée d'une source quasi-périodique et du bruit blanc stationnaire ou du bruit transitoire. Le type du bruit (stationnaire ou transitoire) et le mélange des deux types d'excitation dépendent du type du son. L'effet de la glotte peut être modélisé par un filtre auto-régressif. En ignorant la correspondance en phase, ce filtre est

souvent remplacé par un filtre causal contenant un pole de deuxième ordre. Le conduit vocal peut être bien modélisé par un filtre (AR), ne contenant que des pôles. Le conduit nasal est modélisé par un filtre autorégressif à moyenne ajustée (ARMA) possédant des pôles et des zéros.

L'idée fondamentale est que les échantillons présents peuvent être estimés comme une combinaison linéaire des échantillons passés. Dans un cadre de signal, les coefficients utilisés dans la combinaison linéaire sont obtenus en minimisant l'erreur quadratique de prédiction, sur une trame de signal.

La prédiction à court terme cherche à éliminer la redondance entre les échantillons voisins. Le filtre utilisé est appelé filtre d'analyse LP. Il supprime la structure formantique du signal parole et laisse l'erreur de prédiction de sortie à basse énergie qui est connue sous le nom résiduel ou excitation.

L'inverse du filtre d'analyse est le filtre de synthèse, il modélise le conduit vocal et sa fonction de transfert décrit l'enveloppe spectrale du signal parole [9].

La prédiction linéaire à long terme est utilisée pour exploiter les corrélations à long terme existantes dans les signaux voisés. Ce filtre est appelé prédicteur de pitch. Il exploite la périodicité du signal. L'inverse du prédicteur de pitch est appelé filtre de pitch. Il modélise la fonction de la glotte et sa fonction de transfert décrit la structure harmonique du signal de parole. Le prédicteur de pitch n'a aucun effet pour les signaux non voisés, puisque l'excitation non voisée est aléatoire et son spectre est monotone [10].

I.7.1. Analyse par prédiction linéaire (LPC)

Ce modèle est fortement simplifié et basé sur la production de la parole. L'algorithme de codage par prédiction linéaire (LPC) est l'un des premiers codeurs normalisés qui travaille en bas débit binaire.

La modélisation complète peut-être décomposée en deux parties (figure1.4) :

Une partie synthèse qui effectue un filtrage de fonction de transfert H_{AR} (z). Ce filtre tout-pôle, connu sous le nom de filtre de synthèse LP, permet de reconstruire, à l'aide d'un signal d'excitation approprié, un signal de parole artificiel.

Une partie analyse qui filtre un signal d'entrée avec la fonction de transfert A(z).Ce filtre tout-zéro est défini comme le filtre d'analyse LP et permet d'extraire l'information prédictible

du signal et de définir un signal d'erreur résiduelle entre le signal de parole d'entrée s(n) et son estimation $\hat{s}(n)$ celui-ci s'écrit [11] :

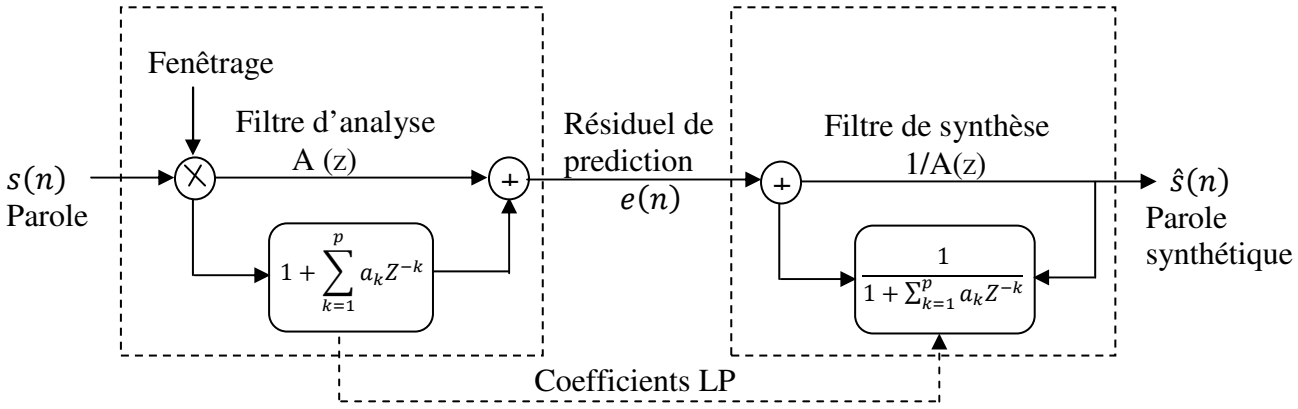


Fig. I.4. Analyse et synthèse LP

$$\hat{s}(n) = -\sum_{i=1}^{P} a_i s(n-i) \tag{1.5}$$

Où a_i, $1 \leq i \leq p$, les coefficient de prediction linéaire (LPC) et P représente l'ordre de prédiction. L'erreur de prédiction est définie par :

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{i=1}^{P} a_i s(n-i) \tag{1.6}$$

Le signal original $s(n)$ peut s'écrire en terme d'un signal prédit et d'un résiduel :

$$s(n) = \hat{s}(n) + e(n) = -\sum_{i=1}^{P} a_i s(n-i) + e(n) \tag{1.7}$$

La transformée en Z de l'equation (1.7) donne :

$$S(z) = -S(z) \sum_{i=1}^{P} a_i Z^{-i} + E(z) \tag{1.8}$$

Ce qui conduit au modèle d'analyse LP

$$E(Z) = S(z).A(z) \quad \text{où} \quad A(z) = 1 + \sum_{i=1}^{P} a_i Z^{-i} \tag{1.9}$$

I.7.2. Prédiction Linéaire A Excitation Multi Impulsionnelle (MP-LPC)

Le codeur MP-LPC (LPC à excitation multi impulsionnelle), elle est un modèle du conduit vocal, pour produire une parole naturelle à bas débit. Cette technique utilise l'enveloppe spectrale à court terme avec un filtre LPC. Elle a été mise au point en 1982 par Atal et Remde dans le but de corriger les limites constatées sur la qualité de la parole obtenue par le codage LPC classique [12]. En effet, le codeur MP-LPC présente l'avantage d'éviter la classification

des sons de la parole en sons voisés et non-voisés. De plus, les erreurs de mesure du pitch inévitables aussi dans la LPC se trouvent ainsi éliminées.

L'algorithme de MP-LPC forme un ordre d'excitation qui se compose d'impulsions multiples non uniforme espacées de façon non uniforme, et approchant au mieux le résiduel de prédiction. Les amplitudes et les positions des impulsions sont déterminées séquentiellement par une procédure d'analyse par synthétise (AbS) qui est à la base de la plupart des codeurs à bas débits. L'algorithme de MP-LPC emploie typiquement 4 à 6 impulsions chaque 5 ms [2].

L'objectif de cette méthode est de permettre une bonne restitution du signal avec un nombre faible d'impulsions. En d'autres termes, on doit chercher $\hat{s}(n)$ qui doit s'approcher de $s(n)$ de façon à avoir $e(n)$ minimale. La figure 1.5 représente le schéma synoptique d'un codeur MP-LPC.

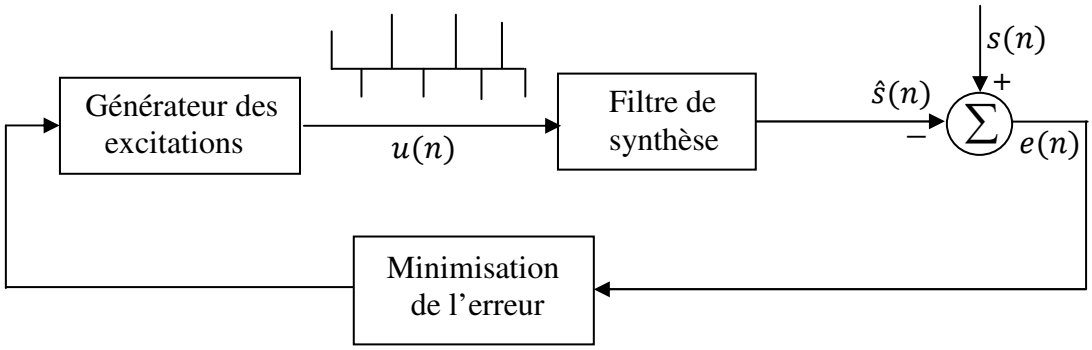


Fig. I.5. Codeur LPC à excitation multi-impulsionnelle ou MP-LPC

L'erreur peut être écrite comme suit :

$$e(n) = s(n) - \hat{s}(n) \tag{1.10}$$

Où $s(n)\ et\ \hat{s}(n)$ sont le signal original et le signal synthétique équivalent.

Faisant appel aux propriétés de l'audition, l'erreur de prédiction $e(n)$ définie en (1.10) sera pondérée et se trouve ainsi " filtrée" par *H(z)* appelé filtre perceptuel et défini par ATAL par la fonction de transfert:

$$H(z) = \frac{A(z)}{A(z/\gamma)} = \frac{1+\sum_{i=1}^{P} a_i z^{-i}}{1+\sum_{i=1}^{P} \gamma^i a_i z^{-i}} \tag{1.11}$$

$A(z)$ est le filtre de prédiction inverse où les a_i sont les coefficients de prédiction, p est l'ordre du filtre et γ est un facteur perceptuel compris entre 0 et 1. Ce facteur γ permet de contrôler la pondération de l'erreur dans les régions formantiques. Pour γ =1, *H(z)* vaut 1 et se trouve sans effet et l'erreur pondérée est égale à la différence entre le signal original et le signal synthétisé.

Pour γ=0 on a *H(z)=A(z).* Dans ce cas, l'erreur est égale à la différence entre le résiduel et la suite d'impulsions placées c'est-à-dire $e(n) - u(n)$. La valeur γ doit être prise entre 0 et 1 suivant la pondération de l'erreur. En pratique, la valeur optimale du facteur γ est déterminée expérimentalement par des tests d'audition. Pour une valeur typique de 8 kHz lui correspond une valeur typique de γ=0.8 [12, 13].

I.7.3. Prédiction Linéaire Excitée par Code (CELP)

Ce codage est une variété des codeurs MP-LPC. Il permet d'obtenir des débits encore plus faibles (<5 kbps). Le codage CELP (Code Excited Linear Prediction) a été introduit par Atal et Schroeder en 1985 [14]. Il fait parti des codeurs hybrides. Depuis, il n'a pas cessé d'être modifié. Aujourd'hui, la majorité des systèmes de codage bas et moyen débit utilisent ce type de codage. Il est très efficace pour les débits intermédiaires de 4.8 kbps à 16 kbps, comme en témoignent les nombreuses normes qui l'utilisent en téléphonie. La Figure I.6 représente le principe du codage CELP [1]. Globalement les codeurs de type CELP modélisent le système de production de la parole en trois étages :

- Un étage d'excitation ;
- Un étage modélisant l'effet des cordes vocales ;
- Un étage modélisant la fonction de transfert du conduit vocal.

Les premières étapes sont les mêmes que pour la LPC. La différence réside dans le choix des fonctions d'excitation. Cette méthode utilise une combinaison linéaire de fonctions stochastiques et périodiques (construction adaptative à partir d'un dictionnaire).

Le système consiste en un filtre de synthèse 1/A (z) à court terme représentant l'enveloppe spectrale du signal. Une prédiction linéaire à long terme représentée par le filtre 1/*B*(z) permet de modéliser la périodicité du signal, c.-à-d le pitch. On parle alors de dictionnaire adaptatif. Sa sortie est ajoutée à un vecteur du dictionnaire innovateur choisi à l'issue de l'analyse par synthèse. La minimisation de l'erreur se fait selon un critère perceptuel, se traduisant par un filtre perceptuel W(z). Ce filtre pondère l'erreur dans le domaine fréquentiel en tenant compte des caractéristiques de l'audition humaine. Il atténue les zones du spectre à forte amplitude (formants) et amplifie les zones de faible amplitude. La forme classique du filtre est la suivante [6, 15]:

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)} \quad ou \;\; 0 < \gamma_2 < \gamma_1 \leq 1 \tag{1.12}$$

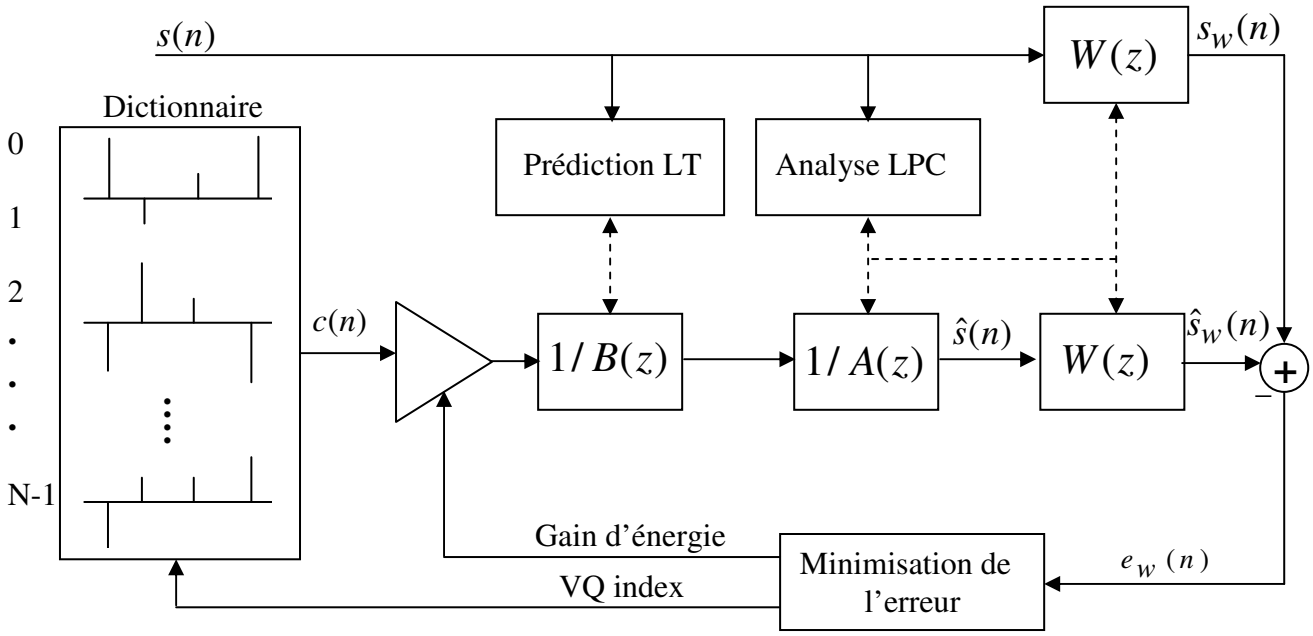


Fig. I.6. Schéma de principe du codeur CELP

Le facteur γ_2 a pour effet de déplacer les pôles vers le centre du cercle par rapport aux pôles du filtre de synthèse $1/A(z)$. Le filtre est donc plus stable et il y a moins de résonance au niveau des formants. Le filtre $W(z)$ favorise alors l'erreur de quantification à avoir une énergie plus importante au niveau des formants hauts en énergie.

Les codeurs CELP travaillent sur des trames consécutives égales en longueurs [6]. Suivant le codeur, la longueur d'une trame varie entre 10ms et 30 ms pour rester dans l'hypothèse de stationnarité du signal de parole. Chaque trame est découpée en sous trames plus courtes (durée typique 5 ms), et sur chaque sous trame on affecte une quantification vectorielle du signal par une technique d'analyse par synthèse.

La quantification vectorielle utilise un dictionnaire de taille $N = 2^k$ séquences de bruit blanc normalisées en énergie. La longueur des ces séquences est égale à une sous-trame. Chaque séquence du dictionnaire est filtrée par le filtre de synthèse $1/A(z)B(z)$ et multipliée par un gain. La sortie obtenue est le signal de parole synthétique qui est comparé au signal original.

I.7.3.1. Prédiction long terme (LTP)

La première caractéristique du codage CELP est l'utilisation de la prédiction à long terme ou LTP. Le signal résiduel obtenu après une prédiction à court terme LPC est un signal quasi périodique pour les sons voisés. L'objectif de la prédiction à long terme est de tirer profit de cette périodicité pour le codage. Un filtre de prédiction à long terme $B(z)$ possède

généralement de 1 à 3 coefficients. Dans le cas d'un seul coefficient, $B(z)$ s'exprime par :

$$B(z) = 1 - bz^{-Q} \tag{1.13}$$

Pour effectuer une prédiction à long terme, il faut calculer le coefficient b (problème linéaire) et la constante Q qui correspond la période T_0 du signal en nombre d'échantillons. Généralement cette dernière est calculée par la méthode d'autocorrelation. Pour des signaux échantillonnés à $F_e = 8$ kHz, la valeur de la période est estimée avec une meilleure précision que la durée d'un échantillon $Te = \frac{1}{Fe} = 125us$. On parle alors de technique de pitch fractionnaire car la valeur de T_0 est estimée à une fraction de T_e près jusqu'à $T_e/6$ [8].

Les codeurs CELP et ses dérivations permettent de réduire le débit jusqu' à environ 4 kbps en utilisant la quantification vectorielle pour coder le résidu. Les normes les plus importantes à mentionner sont la FS1016 à 4.8 kbps [2], le LD-CELP13 (norme G728) à 16 kbps, les codeurs ACELP14 (normes GSM EFR15 à12.2 kbps), le G729 à 8 kbps, le G723.1 à 6.3 et 5.3 kbps, et le G722.2 dans la bande élargie à 13-24 kbps), les codeurs CELP de la norme MPEG (3.85-12.2 kbps dans la bande téléphonique et 10.9-23.8 kbps dans la bande élargie).

Tous ces codeurs de type LPC et utilisent différents signaux d'excitation contenus dans un dictionnaire. Le codeur détermine quelle excitation produit la plus faible erreur et transmet le mot code de cette excitation. Ce mot code remplace l'information voisée ou non voisée. Le codage CELP procure une qualité suffisante pour un débit de 4.8 kbps.

I.7.3.2. Standard Fédéral (FS 1016) à 4.8 kbps

Le fédéral standard FS1016 est un codeur de parole basé sur la technique CELP qui fonctionnant à un débit de 4.8 kbps. Ce dernier a été normalisé conjointement par DoD (Departement of defense) de la défense des USA et les laboratoires de ATAL et Bell en 1991. Depuis sa standardisation, beaucoup de recherches ont été effectuées pour son amélioration [15], qui incorporent les travaux d'amélioration de la qualité et les progrès de codage à bas débit. En général, le codeur CELP de la norme FS1016 [10, 12, 15, 16] découpe le signal de parole d'entrée (échantillonné à 8 kHz) en fenêtres d'analyse de 30 ms (N = 240), dont chacune est divisée encore en quatre sous fenêtres de 7.5 ms. Pour chaque fenêtre d'analyse, le codeur calcule un ensemble de 10 coefficients LPC ($a_i, i = 1, \ldots, 10$) du filtre de synthèse par une prédiction linéaire à court terme. L'analyse LP d'ordre 10 est effectuée par une méthode d'autocorrélation sur chaque fenêtre d'analyse, pondérée par la

fenêtre de Hamming. Pour éviter les problèmes d'estimation de l'enveloppe spectrale, les coefficients LPC sont remplacés par une pondération de la forme $a_i' = a_i\delta^i$, avec $\delta = 0{,}994$. Ce qui conduit à une expansion de 15 Hz de largeur de bande. Les coefficients a_i', qui définissent le filtre de prédiction linéaire $1/A(z)$, sont convertis en coefficients *LSF* (Line Spectral Frequency) qui sont plus adaptés à la transmission. Afin d'éviter les changements rapides dans les coefficients entre deux fenêtres successives, les coefficients *LSF* utilisés dans la détermination de l'excitation des sous fenêtres sont obtenus par interpolation linéaire de deux ensemble de coefficients *LSF* calculés pour deux fenêtres d'analyse successives. Pour chaque sous-fenêtre, le signal d'excitation optimal du filtre de synthèse est déterminé par une modélisation d'ordre 2. Il est donc construit par une combinaison linéaire de deux vecteurs codes, sélectionnés selon une procédure d'analyse par synthèse. Ceci est effectué en minimisant un critère d'erreur perceptuel. Le premier vecteur, d'indice de retard i_a, est extrait d'un dictionnaire adaptatif (prédictif) de 256 vecteurs puis pondéré par un gain g_a. Ce dictionnaire, qui est composé de séquences antérieures d'excitations synthétiques, est utilisé pour modéliser les périodicités à long terme présentes dans le signal de parole voisé. Le second vecteur, d'indice i_s, est extrait d'un dictionnaire stochastique puis pondéré par un gain g_s. Ce dictionnaire qui est de nature statique contient 512 vecteurs codes pseudo aléatoires qui sont quantifiés à 3 niveaux (-1, 0, 1) et distribués selon une gaussienne de moyenne nulle et de variance unité.

En plus de la quantification vectorielle utilisée pour quantifier le signal d'excitation, une quantification scalaire non uniforme est utilisée pour le reste des paramètres. Le standard FS1016 travaille à 4.8 kbps avec des trames de 30 ms de longueur donc l'encodeur doit transmettre au décodeur 144 bits pour chaque trame. La distribution de ces 144 bits est illustrée dans le tableau I.1.

Tableau I.1. Allocation des bits pour le FS1016

Paramètre	Nombre de paramètres par trame	Résolution	Nombre de bit par trame
LPC	10	3, 4, 4, 4, 4, 3, 3, 3, 3, 3	34
Pitch	4	8, 6, 8, 6	28
Gain adaptative	4	5	20
Indice d'excitation	4	9	36
Gain stochastique	4	5	20
Synchronisation	1	1	1
Protection	4	1	4
Expansion	1	1	1
Total			**144**

I.7.4. Prédiction Linéaire à Excitation mixte (MELP)

Le codeur de parole LPC à excitation mixte (MELP) est conçu pour surmonter certaines limitations du LPC10. Il utilise un modèle de production de la parole qui contient des paramètres additionnels pour contenir la dynamique fondamentale du signal avec une précision améliorée. L'idée essentielle est la génération d'un signal d'excitation mélangé comme entrée du filtre de synthèse, où le mixage se rapporte à la combinaison périodiques filtrés d'impulsions et d'un également filtré [17].

L'amélioration apportée par le MELP se situe principalement au niveau de l'étage d'excitation où elle devient "mixte", c'est à dire qu'elle peut prendre plusieurs formes (impulsion, bruit, régime transitoire). Les exigences analogiques recommandées pour le codeur MELP sont pour une bande passante s'étendant de 100 Hz à 3800 Hz, bien que celui-ci opère sur des signaux à bande limitée plus faible.

Un schéma fonctionnel du modèle de MELP de production de la parole est montré sur la Figure I.7. On note que ce schéma est une amélioration de celui du modèle LPC classique [18].

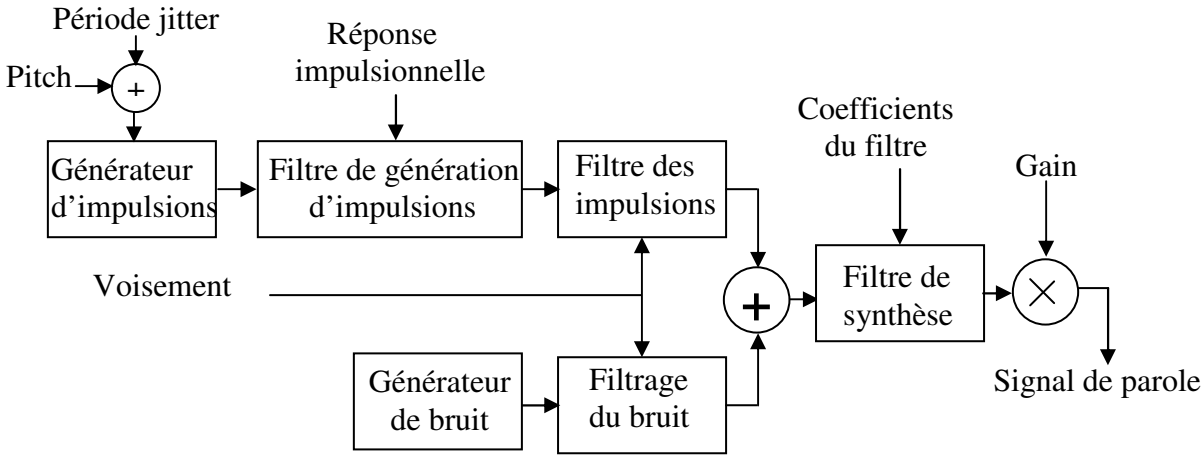


Fig. I.7. Modèle MELP de production de la parole

Le codeur MELP représente la nouvelle norme du DOD à 2.4 kbps. Il remédie donc aux problèmes de la LPC-10 en utilisant une représentation plus souple de l'excitation. Le mélange de la composante impulsionnelle et de la composante de bruit dans l'excitation dépend de la bande de fréquence (5 bandes sont définies entre 0-4000 Hz). La composante impulsionnelle peut être périodique ou apériodique. La dernière joue un rôle dans la modélisation des consonnes plosives non-voisées et des zones de transition entre sons.

Le codeur extrait et transmet les amplitudes des 10 premiers harmoniques du signal résiduel. Cela peut compenser les défauts de la modélisation tout-pôle et améliorer la représentation du signal résiduel. Pour mieux corriger les défauts de la modélisation, le décodeur utilise un filtre adaptatif de renforcement des formants et un autre filtre dont le but est d'étaler l'énergie des impulsions sur une période de pitch. Les paramètres du modèle sont quantifiés de manière efficace, en utilisant des quantificateurs vectoriels pour les LSF et pour les amplitudes des harmoniques. La qualité du codeur surpasse sensiblement celle du codeur LPC-10 [19].

I.8. Evaluation et Perception

Après avoir conçu et réalisé un algorithme de codage, il est nécessaire de mettre ce dernier sous tests afin de l'évaluer.

Dans les communications numériques, la qualité du signal parole est évaluée selon quatre catégories [20]:

- **Qualité diffusion ou broadcast :** qui se réfère aux larges bandes (typiques [50-7000] Hz et [20-20000] Hz pour disques compacts). C'est la plus haute qualité qu'on peut atteindre. Elle nécessite des débits varient entre 32 et 64 kbps.

- **Qualité réseau ou toll :** c'est la qualité qui permet d'entendre la parole sur un réseau téléphonique (pour une bande de 200-3400 Hz avec un rapport signal sur bruit de 30 dB et une distorsion inferieure à 2 ou 3 %).

- **Qualité de communications:** elle implique une certaine dégradation de la qualité de la parole. Néanmoins, elle présente une qualité naturelle hautement intelligible. Cette qualité peut être atteinte à des débits supérieurs à 4 kbps.

- **Qualité synthétique:** la parole synthétique est intelligible, néanmoins, elle n'est pas naturelle et perd la reconnaissance de locuteur.

I.8.1. Mesure de distorsion subjective

Lors d'un test subjectif, on demande à des participants de tester un système de télécommunications dans différentes conditions et de noter, sur une échelle de qualité, la qualité vocale de ce système. Le tableau I.2 représente l'échelle MOS (Mean Opinion Score). D'une manière générale, la qualité dépend de la personne qui la juge. La qualité vocale, dans le cadre des systèmes de télécommunications, est elle aussi dépendante de celui qui l'évalue.

De plus, la perception de la qualité vocale dépend du contexte et de l'environnement dans lesquels est placée la personne qui juge. En effet, si elle est entrain d'écouter un message vocal (contexte d'écoute) ou si elle est impliquée dans une conversation avec un interlocuteur (contexte de conversation), les processus d'attention mis en jeu ne sont pas les mêmes. De même, l'environnement (bruit, l'information visuelle, sonores supplémentaires, etc.), influe sur le jugement de la qualité de la parole [20].

Tableau I.2. Description de l'échelle MOS

Niveau	Qualité de la parole	Niveau de distorsion
5	Excellente	Imperceptible
4	Bonne	A peine perceptible mais pas gênant
3	Moyenne	Perceptible et un peu ennuyeux
2	Pauvre	Ennuyeux mais pas désagréable
1	Mauvaise	Très ennuyeux et désagréable

I.8.2. Mesure de distorsion objective

Les mesures objectives utilisent des fonctions ou des critères mathématiques pour comparer les formes d'onde, les spectres ou les cepstres codés et originaux.

Certaines mesures donnent des informations utiles selon le type de codage testé. Par exemple, le Rapport Signal sur Bruit (SNR : Signal to Noise Ratio) est représentatif pour les codeurs temporels et certains codeurs hybrides, tels que les codeurs de type CELP, qui incorporent des mécanismes de modélisation de forme d'onde. D'autres évaluent certains éléments des algorithmes de codage, tels que les distorsions cepstrales ou spectrales, employées pour calculer la déformation introduite par la quantification des paramètres LPC. Idéalement, les mesures objectives recoupent les résultats obtenus par notre perception subjective de la parole. Toutefois, les tests subjectifs et objectifs peuvent produire des résultats légèrement différents.

I.8.2.1. Mesure de distorsion objective dans le domaine temporel

Rapport signal à bruit (RSB)

Le rapport signal à bruit représente le rapport de la puissance du signal à la puissance du bruit. Le *RSB* s'exprime souvent en décibel selon la relation suivante :

$$RSB = 10\log\left\lfloor\frac{\sum_{n=0}^{N-1} S(n)^2}{\sum_{n=0}^{N-1}(S(n)-\hat{S}(n))^2}\right\rfloor \quad \text{(dB)} \qquad (1.14)$$

Où S(n) représente le signal original, $\hat{S}(n)$ celui décodé et N_s représente la longueur du signal en nombre d'échantillons.

Le *RSB* donne une valeur après avoir traité tout le fichier, donc il n'y a pas moyen de retrouver les instants où les divergences ont été enregistrées. De plus, le *RSB* est dominé par la portion de forte énergie (tranches voisées), alors que le bruit a un effet perceptuel plus important sur les portions de faibles énergies.

- **RSB Segmental (RSBSseg)**

Pour calculer le rapport signal sur bruit segmental RSBseg : on découpe le signal en N_F segments de N_S échantillons chacun et on calcule une moyenne N_F [21] selon.

$$RSBseg = \frac{1}{N_F}\sum_{i=0}^{N_F-1} 10log_{10}\frac{\sum_{j=0}^{N_S-1} S^2(N_S*i+j)}{\sum_{j=0}^{N_S-1}(S(N_S*i+j)-\hat{S}(N_S*i+j))^2} \quad (1.15)$$

Où N_s représente la longueur du signal en nombre d'échantillons.

Le $RSBseg$ est une meilleure mesure que le RSB, mais ce n'est pas toujours le cas quand la trame entière est silencieuse. Pour cela, nous faisons appel à d'autres mesures objectives.

I.8.2.2. Mesure de distorsion objective dans le domaine fréquentiel

La différence entre l'enveloppe spectre du signal parole original et celle du signal codé, qui peut être traduite par une différence entre les fréquences des formants ou entre leurs largeurs, conduit à des sons phonétiquement différents. C'est pourquoi on fait recours à la distorsion spectrale. Une brève description des différentes mesures de distorsion dans le domaine fréquentiel est présentée dans ce qui suit [22]

- **Distorsion d'Itakura-Saito**

La distorsion d'Itakura-Saito, connue sous le nom de mesure de distance du rapport de vraisemblance, mesure le rapport d'énergie entre le signal résiduel obtenu en utilisant le filtre LP avec les coefficients quantifiés et le signal résiduel obtenu en utilisant le filtre LP avec les coefficients non quantifiés. La distance d'Itakura-Saito et donnée par la formule suivante [23].

$$d_{IS} = \frac{1}{2\pi}\int_{-\pi}^{\pi}[e^{V(\omega)} - V(\omega) - 1]d(\omega) \quad (1.16)$$

Avec :

$$V(\omega) = logS(\omega) - log\hat{S}(\omega) \quad (1.17)$$

- **Distorsion spectrale logarithmique**

La mesure de distorsion spectrale logarithmique est la plus fréquemment utilisée, appelée souvent *distorsion spectrale*, Elle est exprimée par :

$$d_{SD}^{P} \frac{2}{2\pi} \int_{-\pi}^{\pi} |10log_{10}S(\omega) - 10log_{10}\hat{S}(\omega)|^{P} d\omega \qquad (1.18)$$

Lorsque *p*=2 la distorsion spectrale est appelée RMS (pour Root Mean Square), elle sera donc définie par :

$$d_{SD} = \sqrt{\frac{1}{\omega_u - \omega_l} \int_{\omega_l}^{\omega_u} \left[10log_{10} \frac{S(\omega)}{\hat{S}(\omega)}\right]^2 d\,\omega} \qquad \text{(dB)} \qquad (1.19)$$

Où:

$S(\omega)$ et $\hat{S}(\omega)$ représentent le spectre du signal original et le signal synthétique.

$$S(\omega) = \frac{G}{|A(e^{j\omega})|^2} = \frac{G}{[1 - \sum_{n=1}^{P} a_n e^{jn\omega}]^2} \qquad (1.20)$$

G est le facteur de gain du filtre LP et $\{a_n\}$ sont les coefficients du filtre.
Lorsque *p*=2 la distorsion spectrale est appelée RMS (pour Root Mean Square). ω_u et ω_l sont des fréquences qui représentent la limite basse et la limite haute de l'intégrale.

- **Distance euclidienne pondérée**

Les $LSFs$ possèdent une relation directe avec la forme de l'enveloppe spectrale. Les formants correspondent aux $LSFs$ voisins (étroitement liés) tandis que les $LSFs$ isolés représentent les vallées. Par conséquent, une distance du carré de l'erreur peut être utilisée pour comparer les vecteurs $LSFs$ originaux et les vecteurs $LSFs$ codés. Soient deux vecteurs $LSFs$ à m dimensions x et $\hat{x}$. La distance euclidienne est donnée par :

$$d(x, \hat{x}) = (x - \hat{x})^T (x - \hat{x}) = ||x - \hat{x}||^2 \qquad (1.21)$$

Afin d'obtenir une bonne estimation de la qualité perceptuelle de l'enveloppe spectrale, on préfère utiliser une distance euclidienne pondérée des $LSFs$:
Où W est une matrice de pondération symétrique et positive de dimensions m x m (m est l'ordre de l'analyse LP). Si W est une matrice diagonale ayant les éléments W_{ii}>0, la distance sera alors :

$$d(x, \hat{x}) = \sum_{i=1}^{m} W_{ii} (x_i - \hat{x}_i)^2 \qquad (1.22)$$

Les mesures dans le domaine perceptuelles sont basées sur les modèles d'audition humaine. Le signal est transformé vers un domaine perceptuel adéquat de telle manière qu'on puisse exploiter les effets de masquage psycho-acoustique. Parmi les mesures perceptuelles les plus utilisées nous pouvons citer: Perceptual evaluation of speech quality (PESQ) et Enhanced Modified Bark Spectrum Distorsion (EMBSD).

L'EMBSD estime la distorsion perceptuelle d'un signal déformé en le comparant au signal original dans le domaine des sons forts (loudness domain) tout en tenant compte du seuil de masquage de bruit modifié et du modèle cognitif basé sur le post-masquage [24].

I.9. Mesure des performances d'un codeur

Pour mesurer les performances d'un codeur, on analyse celui-ci en tenant compte des cinq critères suivants [25]:

- Le débit, qui reflète le degré de compression fourni par l'algorithme de codage;
- La qualité du signal perçu, déterminée par des tests subjectifs basés sur une moyenne des jugements exprimés par un certain nombre d'auditeurs ou par des tests objectifs;
- La complexité algorithmique;
- Le retard de transmission introduit par l'algorithme;
- La robustesse aux erreurs de transmission.

Chapitre II

Transmission de la voix sur IP (VoIP)

II.1. Introduction

Durant les dernières années, le réseau Internet a évolué pour devenir un réseau disponible pour l'information et la transmission de données. En parallèle, un grand nombre de différents services Internet ont été développés, comme la transmission de la voix et de la vidéo, devenus de plus en plus populaires. Chaque utilisateur possédant un ordinateur équipé d'une carte vocale, d'un microphone et d'un logiciel multimédia a la possibilité de se connecter au réseau.

Les communications via le réseau Internet représentent à l'évidence un phénomène en forte croissance, exponentielle depuis plusieurs années, dans le domaine de nouveaux moyens de communication. Internet est devenu très populaire avec l'apparition du Web, il permet de naviguer vers de multiples sites internationaux et de rechercher des informations de toute nature. Il a également favorisé d'autres applications telles que le courrier électronique, les forums de discussion ou le commerce électronique. Dans les années 1995 [26], il est apparu que la transmission de la voix sur Internent pouvait se développer à grande échelle. Avec l'augmentation continue de la vitesse des microprocesseurs et le développement des techniques de traitement de signal, il est devenu réaliste de faire transiter la voix sur IP (Internet Protocol), en lui appliquant le même traitement que les autres types de données circulant sur le réseau. Un réseau informatique n'est pas a priori le support idéal pour assurer le transport de la voix en temps réel, cependant le développement de Web et son faible coût d'utilisation engagent d'ores et déjà de nombreux acteurs sur des applications de téléphonie IP (Internet Protocol), qui dispose d'un potentiel important de nouvelles fonctionnalités. La téléphonie sur Internet apparaît donc comme une des évolutions majeures dans le domaine des télécommunications.

II.2. Architecture TCP/IP pour VoIP

Le réseau Internet global est formé en fait de réseaux d'interconnexion et d'accès appartenant à des opérateurs publics ou privés et de réseaux IP d'utilisateurs. Ces réseaux, qui peuvent être de types différents où qui adoptent des protocoles de communications distincts, sont interconnectés par des routeurs ou passerelles. La passerelle accomplit en effet la

traduction des protocoles pour l'établissement des appels ou des connexions multimédias, en gérant les différents formats rencontrés et en transportant l'information entre les différents réseaux qui lui sont connectés. Outre le transcodage, diverses fonctions sont assurées au niveau des routeurs, notamment la mise en paquets des données, le codage/décodage de la voix ou l'annulation d'échos.

En effet, tous les systèmes acheminant des données n'exécutent pas forcément des protocoles de routage. Ces protocoles créent des tables qui définissent le trajet optimal vers le récepteur, grâce à l'analyse des entêtes des paquets IP, en considérant plusieurs facteurs tels que la durée moyenne de transmission, la charge du réseau ou la longueur totale du message. Le routage consiste seulement à transférer les paquets IP, appelés aussi data-grammes, en fonction des informations contenues dans la table de routage.

II.2.1. Protocoles associés

Un certain nombre de niveaux de protocole permettent d'établir une chaîne complète de communication sur un réseau de type commutation de paquets ou tout système résultant de l'interconnexion de ce type de réseaux. La figure II.1 montre la forme d'un paquet lors d'une transmission.

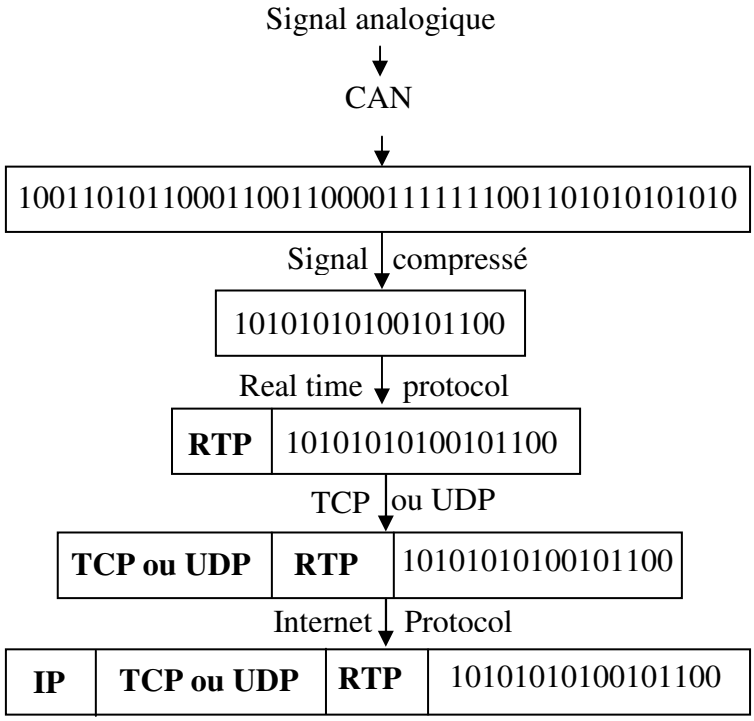


Fig. II.1. Mise en paquet de l'information.

Pour assurer une communication Internet, chaque équipement d'interconnexion impliqué devra posséder le module TCP/IP. Le Protocole Internet, chargé de l'acheminement de

données, encapsule les paquets TCP ou UDP. A l'arrivée dans un routeur, le paquet IP transmis est décapsulé et analysé. Ses informations sont examinées pour savoir vers quel réseau le paquet doit être acheminé et un nouveau paquet IP est constitué selon le réseau désigné puis transmis. Le module d'arrivée extrait le datagramme de niveau supérieur et passe à la couche TCP qui indique les paramètres de service tels que la priorité du paquet ou le codage de sécurité du paquet.

II.2.2. Protocole IP

IP a été inventé par Vinton Cerf et Bob Kahn en 1973 dans le cadre d'un projet de recherche de la Défense américaine (DoD): il s'agissait de trouver des technologies permettant de relier des réseaux transportant des paquets de données [27].

Le protocole IP a été conçu pour permettre les communications entre systèmes informatiques. Sa principale fonction est d'acheminer les paquets IP à travers l'ensemble des réseaux interconnectés, selon l'interprétation d'une adresse identifiant les équipements. Notons qu'une distinction doit être faite entre le nom, l'adresse et le chemin, qui désignent respectivement le terminal recherché, l'endroit où il se trouve et un chemin pour y accéder. Le protocole IP s'occupe essentiellement des adresses, qui sont transportées dans l'entête de chaque datagramme et exploitées par les équipements d'interconnexion pour réaliser le routage. C'est à des protocoles de niveau plus élevé que revient la tâche de lier des noms aux adresses. La tâche qui consiste à transcrire l'adresse en termes de chemin revient, quant à elle, au protocole de bas niveau.

La couche IP dans la suite de protocoles TCP/IP est une couche abstraite où chaque paquet est routé indépendamment, c'est-à-dire qu'un paquet i peut prendre un chemin totalement différent du paquet précédent $i-1$. Un paquet peut donc arriver après celui qu'il précède. Simplement, le protocole IP permet aux paquets de se déplacer sur le réseau Internet indépendamment les uns des autres.

En plus de l'adressage et de l'acheminement des paquets, la couche IP réalise aussi le multiplexage de protocoles ainsi que la destruction des datagrammes ayant transité trop longtemps sur le réseau. Cependant, le protocole IP ne garantit pas la réussite de l'acheminement, le contrôle d'erreur et de flux, ainsi que le séquencement des données.

II.2.3. Protocole H323

Le protocole H.323 figure parmi les plus réputés des protocoles de signalisation pour la téléphonie sur IP. Son nom complet est *Packet-based Multimedia Communications Systems,* ou (Systèmes de communication multimédia fonctionnant en mode paquet). Comme ce nom l'indique, il peut être utilisé pour tous les réseaux à commutation de paquets, en particulier IP. Ce protocole est spécifié pour le traitement de la signalisation en assurant l'établissement, le contrôle et la rupture d'une connexion en temps réel, comme la voix ou la vidéo.

La communauté ITU a normalisé en 1996 le protocole H.323. Ce protocole permet de faire de la téléphonie sur IP. Le H.323 a été réalisé en se basant sur le réseau téléphonique classique. Il est actuellement le plus utilisé par les constructeurs. Le protocole H.323 normalise le type du codage, la procédure de la signalisation et les protocoles temps réels sur IP [27].

II.2.4.Protocol SIP

Session Initiation Protocol (dont le sigle est SIP) est un protocole normalisé et standardisé par l'IETF qui a été conçu pour établir, modifier et terminer des sessions multimédia. Il se charge de l'authentification et la localisation des multiples participants. SIP ne transporte pas les données échangées durant la session comme la voix ou la vidéo. SIP étant indépendant de la transmission des données, tout type de données et de protocoles peut être utilisé pour cet échange. Cependant le protocole RTP (Real-time Transport Protocol) assure le plus souvent les sessions audio et vidéo. SIP remplace progressivement H323.

II.2.5. Protocoles RTP et RTCP

Le protocole RTP (Real-Time Transport Protocol), standardisé en 1996, est adapté aux applications présentant des propriétés temps réel. Il permet ainsi de reconstituer la base de temps des flux (horodatage des paquets et possibilité de resynchronisation des flux par le récepteur), de détecter les pertes de paquets et d'en informer la source ou encore d'identifier le contenu des données pour leur associer un transport sécurisé. Bien qu'autonome, le RTP peut être complété par le RTCP (Real-Time Transport Control Protocol). Ce dernier apporte à la source un retour d'informations sur la transmission et sur les éléments destinataires. Par exemple, un rapport peut regrouper des statistiques concernant la transmission telles que le pourcentage de perte, le nombre cumulé de paquets perdus ou la variation de délai de transmission, appelée gigue. Ces deux protocoles sont adaptés pour la transmission de données en temps réel. Pour le transport de la voix, ils permettent une transmission correcte sur des réseaux bien ciblés tels que les réseaux ATM (Asynchrone Transforme Mode) qui

fournissent une qualité de service adaptée. Des réseaux bien dimensionnés, comme un Intranet, pourront aussi être adéquats. En revanche, les protocoles RTP et RTCP ne permettent pas d'obtenir des transmissions temps réel avec une qualité pour la VoIP. En effet, RTP ne procure pas de réservation de ressources sur le réseau, de fiabilisation des échanges (pas de retransmission automatique et de régulation automatique du débit) ou de garantie dans le délai de livraison (seules les couches de niveau inférieur le peuvent) et de continuité du flux temps réel.

II.2.6. Protocoles TCP et UDP

En principe, les Transport Control Protocol (TCP) et User Datagram Protocol (UDP) doivent pouvoir supporter la transmission de données sur une large gamme de réseaux, depuis les liaisons filaires câblées jusqu'aux réseaux commutés. Les protocoles TCP ou UDP s'intègrent dans une architecture multicouche de protocoles, supportant le fonctionnement de réseaux hétérogènes, supposant que les couches inférieures de communication, sur lesquelles ils s'appuient, leur procurent un service de transmission par paquet. Ces protocoles s'interfacent donc juste au dessus du protocole IP et fournissent un service de transfert de segments de données ou de voix, encapsulés dans un paquet Internet.

Le protocole TCP garantit une fiabilité point à point entre deux processus d'application. C'est un protocole fiable orienté connexion. Il assure les fonctions suivantes [28] :

- Transferts de données de base.
- Contrôle de flux: Utilisation de la fenêtre de contrôle de flux pour réguler les transmissions.
- Connexions: Ensemble composé de "l'adresse IP et du numéro de port de l'émetteur, de l'adresse IP et du numéro de port du récepteur''.
- Remise en ordre des paquets.
- Multiplexage: TCP peut être utilisé par plusieurs processeurs. Ils auront la même adresse IP mais des numéros de ports différents.
- TCP est utilisé pour transporter toute donnée dont l'intégrité et le séquencement sont primordiaux.

Le protocole UDP est un protocole de transport et d'orientation des paquets. Il permet donc aux applications d'échanger des paquets sans accusé de réception ni remise en ordre garantis. Voici les points forts d'UDP [29]:

- Il est efficace pour les diffusions car il ne nécessite pas autant de connexions que de personnes (contrairement à TCP).

- Il est plus rapide, plus simple et plus efficace que TCP, mais au détriment de la fiabilité de la transmission (possibilité de pertes de trame ou de déséquencement).
- Il possède un temps d'exécution court qui permet de tenir compte des contraintes temps réel ou de limitation de mémoire sur un processeur. En conséquence, il est très utilise pour le transport des données temps réel.
- Il peut éventuellement contrôler l'intégrité des données.

II.3. Transmission de la voix

Plusieurs applications de la transmission de la parole sont en constante évolution depuis plusieurs années, tels que les réseaux mobiles (GSM) et les réseaux de type paquet (Internet Protocole, IP). Ces applications reposent sur les techniques de compression et de codage de la parole. Des travaux considérables ont été faits dans ce domaine afin de réduire les débits des codeurs tout en maintenant une bonne qualité de transmission.

Il existe cinq grandes techniques de transmission: la commutation de circuits, le transfert de messages, le transfert de paquets, la commutation de trames et la commutation de cellules. Le transfert est compatible à la fois avec la commutation et le routage tandis que la commutation ne fonctionne qu'en mode commuté.

Dans le cas d'une transmission de la voix, le protocole TCP sera remplacé par UDP, beaucoup plus simple à gérer étant donné qu'il ne fournit pas de contrôle d'erreurs. En effet, le protocole UDP n'opère pas de contrôle de transmission des données, contrairement au TCP, lors d'une communication établie entre deux machines. Il s'agit d'un mode dans lequel la machine émettrice envoie des données sans prévenir la machine réceptrice, qui reçoit les données sans envoyer d'avis de réception à la première. Ce protocole, plus adapté au transport temps réel et donc à la VoIP (Voice over IP), ne garantit alors ni la délivrance du message ni son éventuelle réémission. La plus grande rapidité de restitution de l'information se fera au détriment de la qualité de transmission.

II.4. Généralités sur la voix sur IP (VoIP)

Tout d'abord, il convient de préciser que le terme téléphonique sur Internet ou téléphone IP correspond à la téléphonie utilisant la communication par paquets et les technologies liées à l'Internet, que le réseau soit Internet, Intranet ou Extranet.

Alors qu'il s'agissait simplement au départ d'un complément de l'utilisation d'Internet, la téléphonie sur ce type de réseau semble avoir un intérêt économique évident. La principale raison est propre à la mise en œuvre de la transmission de la voix par paquets qui utilise

mieux les liaisons de télécommunications que la technique de commutation de circuits qui dédie un circuit de bout en bout à chaque communication téléphonique sans tenir compte des temps morts de la conversation. De plus, en téléphonie sur IP (ToIP), une compression de l'information numérique, qui fait passer la voix numérisée d'un débit standard de 64 kbps jusqu'à moins 10 kbps. Elle est pratiquée pour réduire l'occupation spectrale du réseau [27].

Des constructeurs et des opérateurs commencent maintenant à mettre en place des passerelles entre Internet et réseau téléphonique classique et offrent une téléphonique sur Internet concurrente du service habituel. Cependant, la téléphonie sur Internet est encore loin de satisfaire aux exigences de qualité de service attendue. Sans doute est-il trop tôt pour savoir si la téléphonie sur IP peut remplacer le téléphone classique, mais notons que plusieurs études prévoient une importante croissance du marché de la voix sur Internet. En effet, la voix sur IP, qui ne se résume pas à la capacité à établir une connexion vocale entre deux téléphones, peut offrir beaucoup de possibilités en ajoutant de nouveaux services à ceux habituellement fournis par le réseau téléphonique classique. L'idée est d'unifier le transport des informations, voix et /ou données, autour du protocole IP. Etant donné l'engouement actuel pour Internet et les investissements des opérateurs et des fournisseurs d'accès, la téléphonie IP apparaît bien comme une alternative aux réseaux téléphoniques classiques [30].

II.5. Principe de transmission de la voix sur IP

Pour acheminer la voix à travers le réseau IP, il faut réduire au maximum le signal vocal en lui apportant le moins de dégradations possibles car le débit nominal de transport de la voix codée MIC (Modulation par Impulsions et Codage) à travers le RTC (Réseau Téléphonique Commuté) est de 64 kbps alors que la bande passante nominale pour un réseau IP est nettement inférieure à 64 kbps (56, 28.8, 14.4, 8 kbps) [31].

Il existe trois modèles différents de VoIP: La VoIP de PC à PC, la VoIP de PC à téléphone et la VoIP de téléphone à téléphone.

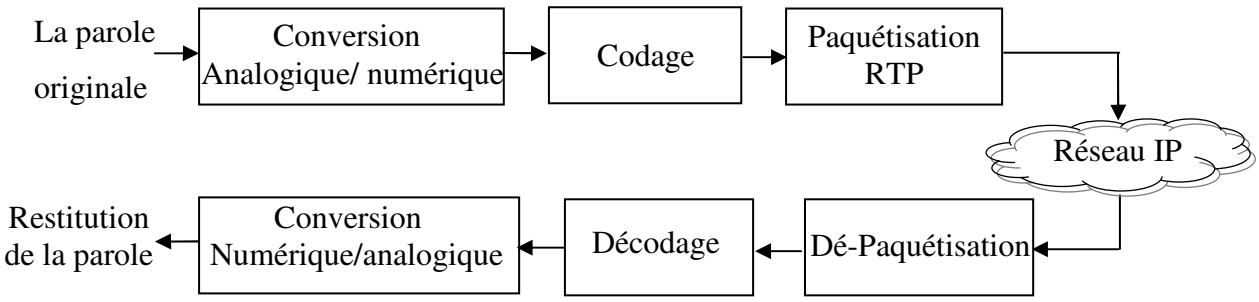


Fig. II.2. Schéma bloc de la transmission de la voix sous le réseau IP

L'établissement d'une connexion VoIP nécessite (figure II. 2)

1. D'abord un CAN qui permet de convertir la voix analogique en signaux numériques.
2. Les bits doivent être ensuite codés en un format adapté à la transmission sous forme de paquets.
3. Il faut ensuite transmettre les données numériques vocales dans des paquets de données à l'aide d'un protocole temps réel (généralement RTP sur UDP sur IP).
4. Il est nécessaire d'utiliser un protocole de signalisation pour appeler les usagers.
5. À la réception, il faut désassembler les paquets, extraire les données, les convertir en signaux analogiques représentant la voix, puis les transmettre pour reconstituer la parole téléphonique.

Pour une communication en VoIP, le signal vocal est numérisé par un codage PCM (Pulse Code Modulation), 8000 échantillons par seconde quantifiés sur 8 bits, puis compressé à l'aide d'algorithmes beaucoup plus élaborés qu'en téléphonie classique. L'information à transmettre est découpée par une procédure de paquétisation, à raison de 20 à 30 millisecondes de parole par paquet, avant l'envoi sur le réseau IP. Les paquets d'informations, qui circulent sur Internet, empruntent des chemins différents et arrivent fréquemment dans le désordre. Les paquets sont alors stockés dans des mémoires tampons, ou buffer, pour être re-séquencés et permettre la décompression de l'information et sa transformation en signal sonore. Les délais de codage et de transit nécessaires à ces opérations étant peu perceptibles par les utilisateurs, les conversations demeurent fluides et sans interruption [27].

II.5.1. Transmission en mode paquet

Après le codage et la division en paquets de l'information binaire au niveau du transmetteur, les paquets contenant la voix sont expédiés à travers le réseau. Les paquets de VoIP interagissent dans le réseau avec les paquets d'autres applications qui sont routés par des connexions partagées vers leur destination.

Le transfert de paquets, passe par deux techniques qui sont la commutation et le routage. Dans le routage, les paquets d'un même client peuvent prendre des routes différentes, tandis que, dans la commutation, tous les paquets d'un même client suivent un chemin déterminé à l'avance.

A l'arrivée, les paquets seront réassemblés et décodés. Le décodage peut être suivi par d'autres étapes aussi, la plus typique étant la compensation de gigue. D'autres exemples sont

la correction d'erreurs et la dissimulation de perte de paquets. Le flux de données numériques est ensuite converti dans une forme analogique et joué sur un dispositif de sortie, typiquement un haut-parleur.

A noter que pour la communication VoIP, qui est bidirectionnelle, la même route existe en direction opposée.

La bande voix qui est un signal électrique analogique utilisant une bande de fréquence de [300 à 3400 Hz], est d'abord échantillonnée numériquement par un convertisseur et codée sur 8 bits. Par la suite, elle est compressée par les codeurs (utilisant des processeurs DSP) selon une certaine norme de compression variable selon les codeurs utilisés. Ensuite, on peut éventuellement supprimer les pauses de silences observés lors d'une conversation, pour ensuite ajouter les en-têtes RTP, UDP et enfin IP. Une fois que la voix est transformée en paquets IP, ces petits paquets IP identifiés et numérotés peuvent transiter sur n'importe quel réseau IP (ADSL, Ethernet, Satellite, switchers, PC, Wifi, etc...).

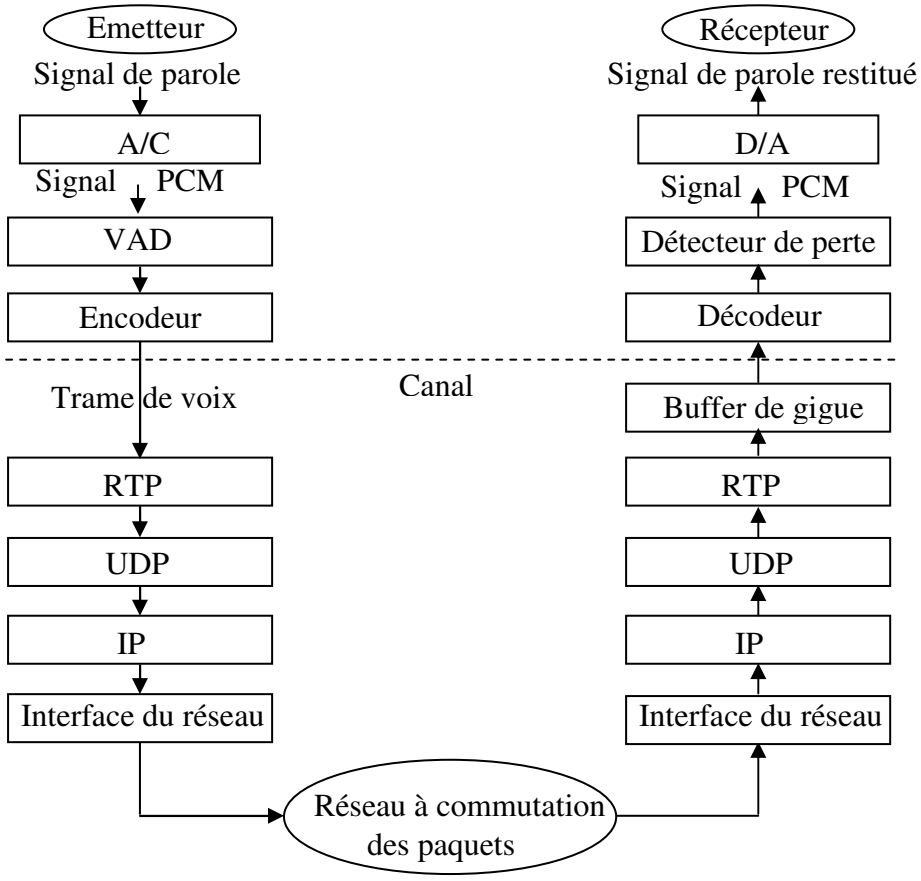


Fig. II.3. Schéma synoptique de transmission de la voix en mode paquets

II.5.2. Les différents codeurs et taux de compression utilisés dans la VoIP

Les principaux codeurs officiels utilisés dans la transmission de la voix sur le réseau IP.

Tableau II.1. Comparative des caractéristiques des Codeurs ITU-T courants.

Codeurs	Débit binaire (kbps)	Délai de codage (ms)	MOS ou Qualité auditive perçue
G.711 PCM	64	0,125	4,1
G.726 ADPCM	32	0,125	3,85
G.728 LD-CELP	15	0,125	3,61
G.729 CS-ACELP	8	10	3,92
G.729a CS-ACELP	8	10	3.7
G.723.1 MP-MLQ	6.3	30	3,9
G.723.1 ACELP	5.3	30	3,65
iLBC Freeware	15.2 13.3	0.125 10	3.9

II.6. Convergence voix données

La téléphonie IP semble être un des média les plus adaptés à la convergence voix-données qui intègre le transport de la voix, de la vidéo et des données sur la même infrastructure. En effet, l'ouverture du système, dont les éléments sont distincts et fonctionnent sur la base de standards, reste un élément essentiel au développement de la téléphonie sur IP. Le véritable enjeu de la voix sur IP dépasse alors le cadre de la téléphonie à grand public. La possibilité de combiner différents types d'informations sur le même réseau de transport permet d'imaginer de nouvelles applications telles que la vidéoconférence sur Internet ou la messagerie unifiée pour envoyer indifféremment des messages électroniques, de voix et des télécopies via l'Internet. [27]

La convergence voix-données étant accomplie à travers le protocole IP supporté par tous les réseaux existants, la téléphonie IP ne requiert alors pas beaucoup d'investissements. Dans un système de communication traditionnel, le réseau local Internet (LAN : Local Area Network) et le réseau téléphonique sont entièrement distincts alors que dans un système convergent voix données, la commutation LAN se situe au cœur de l'architecture. Dans ce cas les routeurs, les passerelles, les serveurs des messageries ou de téléphonie sont interconnectés. Ceci permet à l'ensemble des éléments qui composent l'architecture convergente d'échanger des informations. L'homogénéisation des réseaux est donc aussi à mettre en avant. La convergence permet en effet l'élimination de coûts multiples liés à l'infrastructure, à

l'administration et à la maintenance. Elle réduit, de ce fait, les coûts des réseaux dédiés et d'expansion du système.

II.7. Applications liées à la voix sur IP

Le système vocal est complexe. Il est basé sur des ondes sonores de fréquences différentes. Le spectre des fréquences perçues par l'oreille humaine s'étale de 100 Hz à 20 kHz. Cette fourchette est, cependant, à réduire si l'on veut distinguer les fréquences utiles des fréquences audibles. En effet, la quasi-totalité d'un message sonore est compréhensible dans la fourchette [300-3400] Hz. Cette dernière correspond, d'ailleurs, à celle utilisée par le téléphone standard.

Plusieurs types d'applications, liées à la transmission VoIP, sont énumérés ici [33, 34] :

- En plus des applications précédemment citées, qui émergent actuellement, on peut citer l'utilisation de centres d'appels téléphoniques dans le cadre de transaction de commerce électronique sur Internet. Même avec une qualité moyenne de communication, le client appréciera à l'évidence d'obtenir des renseignements en temps réel, sans manœuvre supplémentaire, pour passer à la commande.

- D'autres applications mixant l'audio, la vidéo et les données peuvent être imaginées telles que l'utilisation d'un son de haute qualité en téléconférence, l'intégration du son dans des jeux réseau, etc.…

- Une autre offre possible est la téléphonie à grande distance et à faible coût quitte à renoncer à la qualité de service offerte par les opérateurs de télécommunications classiques.

- Le dernier domaine d'application est purement orienté vers le téléphone et dont le développement sera très lié à plusieurs facteurs : la généralisation d'offres de passerelles entre réseaux locaux téléphoniques et Internet, l'amélioration da la qualité de service, les prix pratiqués, etc.…

II.8. Qualité de service (QoS) dans le réseau IP

La qualité de service est une condition nécessaire au passage du multimédia dans les réseaux IP. Les réflexions menées sur l'architecture TCP/IP pour aller dans ce sens sont nombreuses.

La qualité de service (QoS) est une notion née chez les opérateurs de télécommunication vers 1997 [34]. Elle désigne l'aptitude à pouvoir garantir un niveau acceptable de perte de paquets, défini contractuellement, pour une utilisation donnée (voix sur IP, vidéoconférence,

etc.). D'autre part, la qualité de service est un ensemble de contraintes que le réseau doit respecter pour offrir un niveau de service approprié à la transmission avec une bonne perception dans un délai raisonnable.

Les paramètres de base caractérisant la QoS d'un trafic IP comprennent le débit, le délai, la gigue et le taux de perte de paquets. La maîtrise du délai de transmission est essentiel pour bénéficier d'un véritable mode conversationnel et minimiser la perception d'écho et éviter les pertes de paquet qui influent sur la qualité perçu [35].

Les paramètres de performance QoS liés au transfert d'information peuvent être définis dans le tableau II.2.

Tableau. II.2. Critères de performance et les paramètres de QoS

Critères de performance	Paramètres de QoS
Vitesse	Débit, délai
Exactitude	Probabilité d'erreur
Fiabilité	Probabilité de perte

La mise en œuvre d'une solution de voix sur IP au niveau local ne pose pas beaucoup de problèmes de qualité de service, vus les hauts débits disponibles sur des interfaces LAN (Ethernet, Fast Ethernet, Giga bits,….). Néanmoins, toutes les préconisations des constructeurs recommandent la mise en œuvre de la gestion de la qualité de service, y compris sur les LANs. En effet, même sur un LAN haut débit, des phénomènes de congestion peuvent survenir et produire des variations dans le délai de transmission de paquets de la voix (micro-coupures dégradant la qualité auditive perçues par les correspondants). Ces congestions surviennent sur les interfaces de concentration des flux.

II.9. Les contraintes de la voix sur IP

La transmission de la voix sur réseau IP se fait par transmission de paquets. Ce protocole peut être sujet à des congestions, auquel cas des paquets seront éliminés aléatoirement. Au niveau du récepteur, ces paquets manqueront. Ceci provoque des pertes de trames du signal de parole transmis. Evidemment, cette perte dégrade la qualité de la voix au niveau du récepteur. Les principales causes de ces pertes associées à la VoIP sont : le délai, la gigue et le taux de perte des paquets [28, 33]. La figue II.4 schématise ces causes.

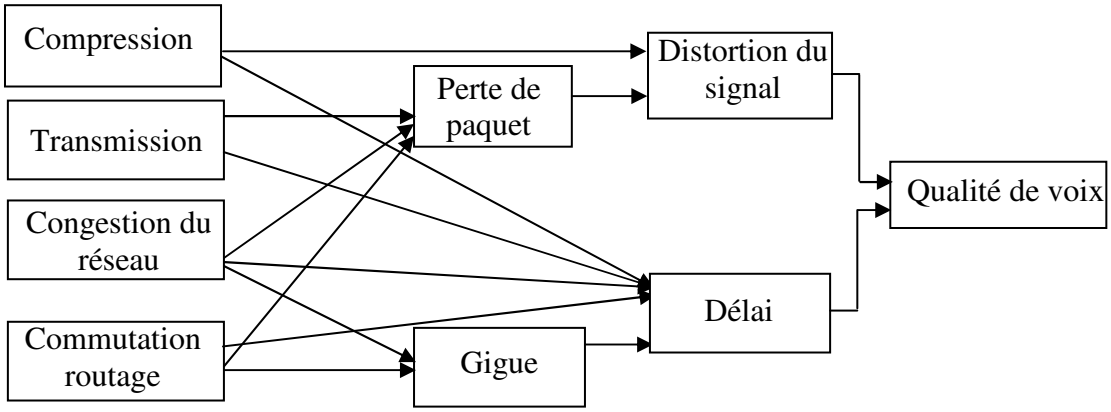


Fig. II.4. Les contraintes de la VoIP

II.9.1. Délai de transfert

Le délai de transfert du paquet IP est défini comme la différence entre le moment de réception et celui de transmission du paquet et doit être inférieur à T_{MAX}. Pour une communication optimale, le délai de transmission total doit être inférieur à 150 ms [36].

En téléphonie, la maîtrise du délai de transmission est un élément essentiel pour bénéficier d'un véritable mode conversationnel et minimiser la perception d'écho.

Les délais ou retards observés lors d'une communication de la voix sur IP (figure II.5) sont dûs à quatre types de problèmes [37] :

- Le retard dû au couple codeur- décodeur. Celui-ci a généralement besoin d'un délai d'une trame pour fonctionner, mais ce retard peut coïncider avec celui du système d'exploitation. Cependant, certains codeurs ont besoin de connaître quelques échantillons en avance pour éventuellement anticiper une perte de trames. Enfin, certains traitements plus complexes au sein des codeurs-décodeurs comme une transformée de Fourier peuvent rajouter l'équivalent d'une trame de retard.

- La gigue ou jitter : du fait que les files d'attente des routeurs sont plus ou moins remplies de manière aléatoire, les paquets IP sont susceptibles d'arriver à destination avec des délais différents. Ils ne peuvent donc être utilisés immédiatement sous peine de laisser des silences entre les trames. Afin d'éviter de trop dégrader la qualité, des buffers de gigue sont mis en place afin de resynchroniser les trames. Cela introduit cependant un retard mais qui est jugé moins gênant que la perte de la trame [38]. Ces buffers permettent aussi de remettre les paquets dans le bon ordre afin d'éviter le dé-séquencement si les paquets suivent des chemins différents.

• L'influence du système d'exploitation et sa capacité à aller récupérer les données mises dans le buffer de la carte son lorsque celui-ci est plein. Selon le système d'exploitation, les données mémorisées dans le buffer de la carte son seront disponibles plus ou moins rapidement, entraînant éventuellement un retard.

• Le nombre de trames audio contenues dans un paquet IP. Plus ce nombre est important, plus le délai est important puisque l'on attend un certain temps avant d'envoyer les données. Ceci a néanmoins l'avantage de diminuer le débit du fait de l'économie d'entêtes IP.

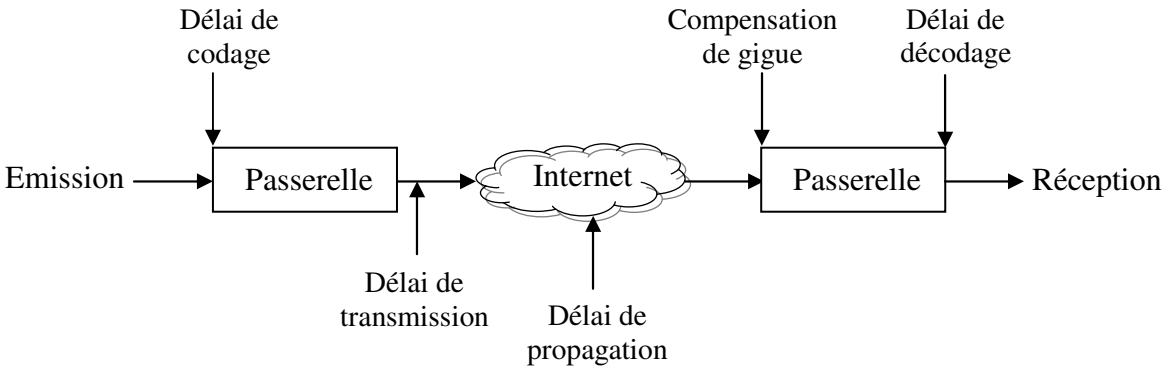


Fig. II.5. Délais causés lors d'une transmission par paquet

Les chiffres du tableau II.3 (tirés de la recommandation UIT-T G114) sont donnés à titre indicatif pour préciser les classes de qualité et d'interactivité en fonction du retard de transmission dans une conversation téléphonique. Ces chiffres concernent le délai total de traitement, y compris le temps de transmission de l'information sur le réseau.

Tableau. II.3. Délais requis pour la VoIP en fonction de la classe d'appartenance

Classe n°	Délai	Commentaire
1	0 à 150 ms	Acceptable pour la plupart des conversations.
2	150 à 300 ms	Acceptable pour des communications faiblement interactives.
3	300 à 700 ms	Devient pratiquement une communication semi duplex.
4	Au delà de 700 ms	Inutilisables même pour une communication semi duplex.

II.9.2. Pertes de paquets

On définit un paquet perdu comme étant un paquet qui n'arrive pas dans un temps ou dans un délai bien déterminé. Elle entraîne la disparition d'un ou plusieurs échantillons du flux

voix, on parle alors de distorsions du signal. Selon le nombre de paquets perdus, la qualité sonore en bout de ligne peut s'en ressentir.

On estime les pertes des paquets dans le réseau IP, par deux types :

- Taux de perte : Il correspond au nombre total des paquets perdus par rapport au nombre total des paquets transmis.
- Taux de paquet faux : est le nombre total de paquets faux observés dans un intervalle ΔT divisé par T.

- Si aucun mécanisme performant de récupération des paquets perdus n'est pas mis en place, alors la perte de paquets IP se traduit par des ruptures au niveau de la conversation et une impression de hachure de la parole. Cette dégradation est bien sûr accentuée si chaque paquet contient un long temps de parole (plusieurs trames de voix). Par ailleurs, les codeurs à très faible débit sont généralement plus sensibles à la perte d'information, et mettent plus de temps à reconstruire un codage fidèle.

Il est à noter qu'il existe des algorithmes de correction de perte de trames intégrés dans des codeurs. Il convient de distinguer deux méthodes importantes souvent retenues dans ce domaine: la correction de la perte de paquets (PLC - Packet Loss Concealment) et la correction d'erreur par anticipation (FEC – Forward Error Correction) [28].

- Pour la PLC, aucune donnée supplémentaire n'est envoyée en anticipation au récepteur qui doit donc faire au mieux. La solution par défaut est l'insertion d'un silence de la longueur de la trame perdue. Une autre solution, meilleure en termes de qualité audio, consiste á remplacer le phonème perdu par un bruit, une tonale ou par lui-même suivant que le son est respectivement voisé ou non-voisé ou seulement une partie de ce phonème. La dernière possibilité est l'interpolation en fonction de la trame précédente. L'avantage majeur de la PLC est qu'elle n'introduit pas de retard supplémentaire.

- Pour la FEC, des données supplémentaires redondantes sont transmises au sein d'une trame par l'émetteur, concernant la trame précédente ou la suivante. Celles-ci sont utilisées par le récepteur en cas de perte de trames. Cela aboutit généralement à la restitution d'une bonne qualité audio par rapport à la PLC, mais cela introduit un délai supplémentaire et une augmentation de débit. Son utilisation principale est généralement la diffusion (streaming) qui a moins de contraintes de délai par rapport à la VoIP.

II.9.3. Erreurs binaires

Elles se produisent notamment sur des réseaux de données avec ou sans fil. Elles peuvent aboutir à une suppression du paquet IP lorsqu'une erreur se produit.

II.9.4. L'écho

L'écho survient à la suite d'une transmission de signaux couplés à une voie de retour à leurs sources. Le locuteur entend sa propre voix avec un décalage temporel dû à la transmission. Il résulte du passage de la transmission. Il s'agit d'un phénomène qui gêne la communication vocale. Le signal se produit avec retard notable.

II.10. Avantages de la voix sur IP

Malgré les inconvénients cités ci-dessus, la voix IP offre plusieurs avantages par rapports au réseau classique Les avantages les plus marqués sont les suivants [28, 39]:

- Intégration de la voix, des données et de la vidéo en un seul réseau.
- Optimisation et efficacité d'exploitation de la bande passante.
- Diminution des tarifs et des coûts des communications.
- Facilité d'administration, de supervision et de maintenance.
- Simplification de la gestion des trois réseaux (données, voix et vidéo) par ce seul transport.
- Choix d'architectures : il existe de nombreuses architectures possibles contrairement au réseau RTC. Cela permet de s'adapter aux besoins définis pour le service proposé et aux limites du matériel utilisé.
- Augmentation et amélioration de la qualité des services.
- Evolution vers un réseau de téléphonie sur IP parce que celle-ci repose totalement sur un transport VoIP.

Conclusion

Dans ce chapitre nous avons présenté un aperçu global sur la transmission de la voix sur le réseau IP, les éléments constituant ce réseau et les différents protocoles utilisées dans la téléphonie IP. Nous avons aussi abordé la qualité de la voix et les facteurs affectant la qualité de service.

Chapitre III

Codage par Descriptions Multiples

III.1. Introduction

Les réseaux IP se montrent robustes et flexibles aux pertes qui se produisant sur des données de type fax, images ou bande-son. En effet, les données perdues peuvent être réémises à partir de la source et réintroduites à la destination sans que l'intégralité ou la qualité des fichiers reçus ne soit altérée. Cependant, dans des applications en temps réel, telles que la VoIP, la procédure de réémission risque d'être impraticable tant les exigences temporelles sont strictes. Ainsi, lors de l'envoi de séquences de paroles, c'est le protocole UDP/IP qui est utilisé. En effet, comme vu plus haut et contrairement au TCP/IP, ce protocole ne requiert pas d'interactions initiales avec le destinataire, ni de réémission de paquets lors de pertes. Des techniques de dissimulation ont été employées soit au niveau de l'émetteur ou au niveau du récepteur afin de combattre les erreurs et la perte des paquets.

Parmi ces dernières, nous citons la technique de codage par description multiple qui sera utilisée dans notre travail.

III.2. Définition

Le codage par descriptions multiples (MDC) trouve son origine dans les laboratoires Bell, dans les années 1970 [40]. Le problème qui se posait à l'époque était d'assurer la fiabilité des communications téléphoniques. La transmission sur deux lignes distinctes pouvait permettre de maintenir la continuité d'une conversation en cas de panne d'une des deux lignes. Il devenait intéressant dans ce cas de ne pas dupliquer l'information, mais plutôt de la repartir entre les deux lignes de manière à ce que la qualité soit maximale lorsque les deux lignes fonctionnent et se dégradent mais, gardent une qualité pour poursuivre la conversation en cas de panne de l'une des deux lignes. Cela a donné une première définition du codage par descriptions multiples. Seuls des résultats théoriques ont été publiés sur le sujet au début des années 1980 [41]. Récemment, le codage par descriptions multiples a connu un regain d'intérêt. De nombreuses techniques permettant de mettre en œuvre ce principe ont été proposées. L'application principale du codage par descriptions multiple est la transmission de

données soumises à des contraintes de délai et de perte, telles que la parole et la vidéo sur des réseaux de type Internet, filaires ou non filaires.

Le codage par description multiple est une technique intéressante pour lutter contre les pertes et les erreurs de transmission. En MDC, la source est codée en plusieurs flux appelés descriptions. Il s'agit en fait de créer plusieurs représentations distinctes mais corrélées d'une source qui seront transmises sur des chemins différents. La réception d'une description quelconque doit permettre une reconstruction de la source avec un niveau de qualité acceptable.

La reconstruction et la qualité s'améliorent avec le nombre de descriptions reçues. Chaque réception de description supplémentaire doit permettre d'améliorer la qualité de reconstruction. La qualité optimale est obtenue quand toutes les descriptions sont reçues.

L'idéal serait donc d'avoir un système de codage qui repartit les données dans des paquets et dont la qualité de reconstruction dépend uniquement du nombre de paquets reçus, et non pas quels paquets sont reçus. On dit dans ce cas, que les paquets se raffinent mutuellement.

Le cas particulier du codage à deux descriptions a fait l'objet d'études approfondies, aussi bien théoriquement que pratiquement. Par construction, le codage par descriptions multiples est bien adapté à la transmission sur plusieurs canaux indépendants ou sur un canal à effacements sans mémoire. Il a également l'avantage de favoriser le respect des contraintes de délai, puisqu'il n'y a pas besoin d'attente que la totalité des descriptions soient reçues pour pouvoir décoder les données [42].

III.3. Codage par description multiple à deux descriptions

Soit la source $\{X_K\}$ transmise sur deux canaux distincts. Les canaux ont deux états de fonctionnement possibles : soit ils sont en état de marche et transmettent sans erreur tous les symboles, soit ils sont défaillants et ne transmettent aucun symbole. Le récepteur connait l'état des canaux mais pas le transmetteur qui envoie systématiquement deux descriptions sur les deux canaux.

Au récepteur, trois décodeurs correspondant à trois situations différentes sont utilisés et le cas intermédiaire. Le décodeur central est utilisé quand les deux canaux sont en état de fonctionnement. Les deux décodeurs latéraux sont utilisés quand seul le canal qui leur est associe fonctionne. On suppose qu'il existe toujours au moins un canal en état de

fonctionnement. Le débit du canal i est noté R_i. En notant $\{\hat{X}_k^{(i)}\}$ la séquence reconstruite par le décodeur i, Les distorsions D_0, D_1 et D_2 s'expriment pour chaque récepteur en utilisant l'expression suivante :

$$D_i = \frac{1}{N}\sum_{k=1}^{N} E[\delta_i\left(X_k, \hat{X}_k^{(i)}\right)], \qquad i \in \{0,1,2\} \tag{3.1}$$

Où δ_i est une fonction à valeurs réelles non négatives :

$$\delta_i(x, \hat{x}) = (x - \hat{x})^2 \tag{3.2}$$

La fonction distorsion-débit pour une source gaussienne est donnée par :

$$R(D) = \frac{1}{2} log \frac{\sigma_x^2}{D} \tag{3.3}$$

Où σ_x^2 est la variance de X. On peut noter que $R(D)$ donne l'information mutuelle minimale nécessaire pour reproduire la source X avec la distorsion moyenne D. Si on suppose que σ_x^2 est égale à 1, cette relation est inversible et on peut écrire:

$$D(R) = \exp(-2R) \tag{3.4}$$

Le cas traditionnel implique deux descriptions comme montrées dans figure. III.1. Le débit binaire total R_T, est tel $R_T = R_0 + R_1$ et correspond à la somme des débits des deux descriptions. Autrement dit, les distorsions observées au récepteur dépendent de la description parvenant au récepteur. Si les deux descriptions sont reçues, la distorsion résultante (D_c) est plus petite que celle correspondant au cas ou seule une description simple est reçue (D_0 ou D_1). D'une façon générale, nous nous référons à D_i, i= 0, 1, les distorsions latérales et par D_c à celle centrale.

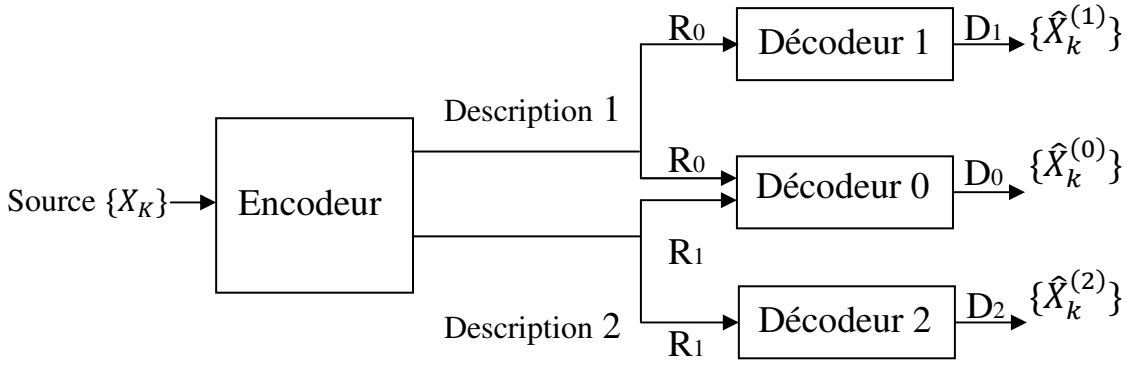


Fig. III.1. Exemple de codage en MD de deux canaux

Le but de la MDC est de construire de manière optimale la partie centrale et les deux autres descriptions. D_0, D_1 et D_2 sont les moyennes des distorsions associées respectivement à la partie centrale de décodage et aux deux côtés des décodeurs. En supposant que chaque description est acheminée par l'intermédiaire d'un canal indépendant symétrique avec ρ la probabilité de perte de paquets, la distorsion totale du système MDC sera donnée par l'expression suivante [43]:

$$D_{MDC} = (1-\rho)^2 D_0 + \rho(1-\rho)(D_1 + D_2) \quad (3.5)$$

Les deux descriptions sont dites «équilibrés» lorsqu' elles possèdent le même taux ($D_1 \approx D_2$). On peut classer les techniques MDC en trois catégories, selon la façon dont les corrélations entre les descriptions sont saisies: la MDC à base d'un indice de vecteur où d'une matrice, la MDC à base de treillis et la MDC basée sur des transformations.

La différence entre les distorsions centrale et latérales devient de moins en moins pertinente puisque le but ultime est de réduire au minimum la distorsion entre la source d'origine à l'émetteur et sa reproduction au niveau du récepteur.

III.4. Codage par description multiple à N descriptions

Dans un premier temps, l'étude théorique du codage par descriptions multiples a été limitée au cas particulier du codage à deux descriptions. Récemment, Puri et Pradhan [44] sont parvenus à caractériser les performances de la MDC utilisant N descriptions avec $2^N - 1$ décodeurs. Ce résultat est une généralisation obtenue du cas N=2. Plus précisément, ces auteurs considèrent le cas où la source a un débit entropique fini et où la communication s'effectue sans pertes quelque soit le nombre de descriptions reçues. En normalisant le débit de la source et en supposant une utilisation identique pour chaque canal, ce dernier doit alors recevoir un débit de 1/(N –k), ou k est le nombre de descriptions reçues [45].

Le schéma de la figure III.2 représente le cas général de codage par description multiple à N descriptions.

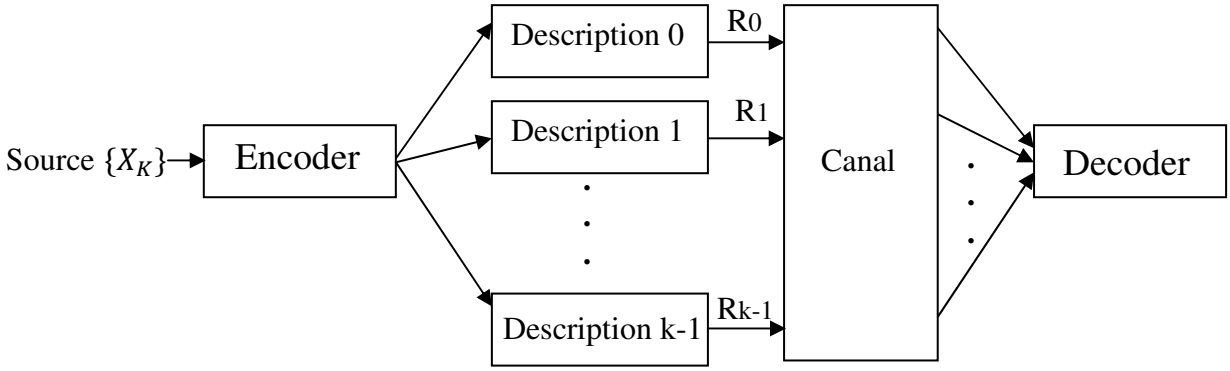


Fig. III.2. Schéma général de codage en MD à k descriptions

Considérons le cas N = 3 et que l'on peut généraliser par la suite au cas N>3. Soit une source X ayant pour densité de probabilité la fonction q(x). X est segmentée en 3 couches Y_1, Y_2 et Y_3. Soient y_3 l'alphabet de Y_3 et y_{ij} l'alphabet de Y_{ij} où i ∈ I I= {1, 2} et j ∈ J , J= {1, 2, 3}. Le débit total est partitionné en trois débits : $R_T = R_1 + R_2 + R_3$. La couche de base est codée par un code correcteur d'erreurs (3, 1) en trois variables Y_{11}, Y_{12}, Y_{13} pour un débit R_1 comme le montre la figure III.3. La réception de deux descriptions doit être suffisante pour décoder la deuxième couche. La réception de deux descriptions permet de recevoir une des paires (Y_{11},Y_{12}), (Y_{12},Y_{13}) ou (Y_{13},Y_{11}). Chacune de ces paires peut être traitées comme information adjacente pour estimer X. Grâce au codage de source avec information adjacente, une seule couche de raffinement est alors nécessaire. C'est cette information qui doit être décodée lors de la réception de 2 paquets. L'information adjacente pour le décodage de Y_{21} est donc constituée de Y_{22} et Y_{23}, mais également de Y_{12}, Y_{13} et Y_{11}. La dernière couche, Y_1, qu'il est possible de décoder seulement si tous les paquets ont été reçus, peut être envoyée à un débit R_3 par paquet en utilisant une approche de raffinement progressif.

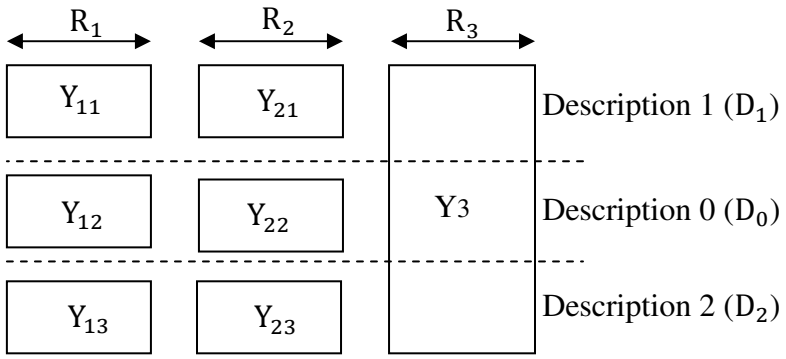


Fig. III.3. Organisation du codage MD à trois descriptions.

III.5. Codage par Description Multiple Basé sur la Quantification

En 1993, Vaishampayan [46] a introduit la quantification scalaire à descriptions multiples (multiple description scalar quantization, MDSQ), ainsi que la quantification vectorielle à descriptions multiples (multiple description vector quantization, MDVQ).

III.5.1 Description Multiple à quantification scalaire (MDSQ)

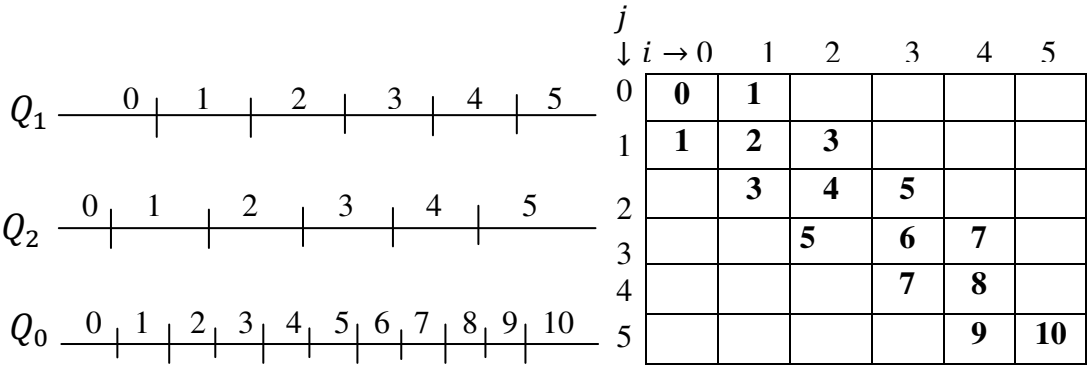
Un quantificateur scalaire à description multiple (MDSQ) est un quantificateur scalaire conçu pour le codage de deux descriptions [46]. Le MDSQ se compose d'un quantificateur scalaire (au sens classique), combiné à un procédé de codage qui assigne, chaque niveau de quantification à une paire d'index qui sépare l'information de chaque échantillon quantifié en deux descriptions complémentaires du même échantillon. Un tel système transmet l'information complémentaire sur deux canaux. Le but est de pouvoir reconstruire le signal avec la fidélité la plus élevée quand les deux descriptions sont disponibles ou bien reçues au décodeur. Un exemple est illustre sur la figure III.4.

Dans MDSQ, les corrélations entre les deux descriptions sont représentées par l'intermédiaire de l'indexation de la matrice qui lie les deux codebooks latéraux de description avec le codebook central de description [47].

Une exécution simple de MDSQ emploie deux quantificateurs dans les régions de décision. Ces deux quantificateurs sont appliqués au même signal, chacun rapporte une description.

La MDSQ est utilisée essentiellement pour le codage de deux descriptions, est constituée de deux parties: un quantificateur scalaire (au sens classique), et une assignation d'index qui sépare l'information de chaque échantillon quantifié en deux descriptions complémentaires du même échantillon.

Considérons le cas d'une transmission sur deux canaux, comme montrée sur la figure III.4. La quantification scalaire par descriptions multiples à (M_1;M_2) niveaux de reconstruction va projeter les échantillons du signal x sur les niveaux de reconstruction.
Les deux codeurs imposent une partition A = $\{A_{ij}, (i,j) \in \mathbb{C}\}$ de l'ensemble R, telle que $A_{ij} = \{x/Q_1(x) = i, Q_2(x) = j\}$. Cette partition est définie arbitrairement par une table d'index, telle que celle de la figure (III.4.b). L'allocation des index doit être telle que si les deux index i et j sont reçus, la qualité de reconstruction du signal est équivalente à une quantification fine, et si seul l'un des deux index est reçu, le niveau de reconstruction est équivalent à une quantification à pas "grossier".



j / $i \rightarrow$

	0	1	2	3	4	5
0	**0**	**1**				
1	**1**	**2**	**3**			
2		**3**	**4**	**5**		
3			**5**	**6**	**7**	
4				**7**	**8**	
5					**9**	**10**

Fig. III.4. (a) Quantification décalée, (b) Représentation matricielle

Dans cet exemple les indices des quantificateurs correspondent aux colonnes et aux lignes de la matrice

Ainsi, si chaque quantificateur a un débit de R bits par échantillons de signal, l'erreur de reconstruction est équivalente à l'erreur qui aurait été obtenue avec un quantificateur unique avec un débit de R+1 bits, lorsque les deux descriptions sont reçues. Si seule l'une des deux descriptions est reçue, alors l'erreur de reconstruction est équivalente à l'erreur obtenue avec un quantificateur induit un débit de R bits. Donc, avec un tel quantificateur MDSQ, en l'absence de pertes, il est nécessaire de transmettre 2R bits pour obtenir des performances équivalentes à un quantificateur avec un débit de R+1 bits.

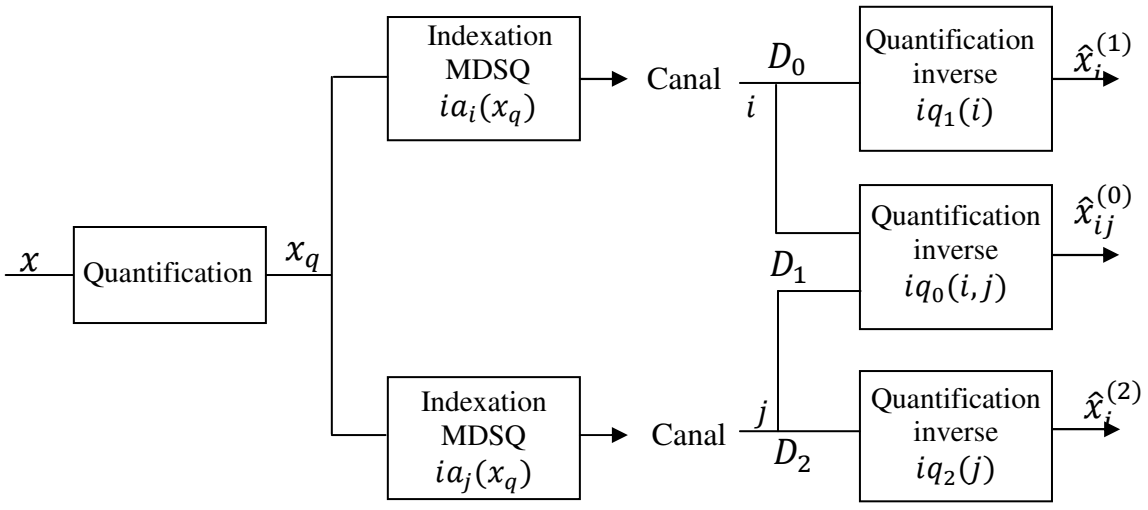


Fig. III.5. Quantification scalaire par descriptions multiples

La distorsion totale dans ce cas est donnée par l'expression suivante :

$$D_{total} = D_0 + \frac{\rho}{1-\rho}(D_1 + D_2) \tag{3.6}$$

III.5.2 Description Multiple à Quantification Vectorielle (MDVQ)

La MDSQ peut s'étendre à la quantification vectorielle. On parle alors de quantification vectorielle à descriptions multiples (MDVQ) ou de quantification vectorielle sur réseau de points réguliers à descriptions multiples (MDLVQ) pour (Multiple description lattice vector quantization). Que ce soit pour le codeur ou bien pour les trois décodeurs, l'espace de représentation d'un vecteur de longueur N est R^N. Le principe de la MDLVQ est similaire à celui de la MDSQ. L'idée consiste à étiqueter chaque point du réseau régulier par une paire d'indices correspondant respectivement à chacune des deux descriptions.

Toute la difficulté de la MDLVQ réside dans cet étiquetage. En effet, comme pour la MDSQ, le nombre de combinaisons est trop important pour envisager une recherche exhaustive. De plus les heuristiques utilisées pour la MDSQ reposent sur un ordre de balayage, qui ne peut pas être défini dans un espace de dimension N.

Si par exemple la paire ‘‘ab’’ a été codée, la réception du premier indice entrainera le décodage du point ‘‘a’’, alors que la réception du second indice entrainera le décodage du point “b’’.

III.6. Description Multiple à Quantification Codée en Treillis (TCQ)

La quantification codée en treillis (TCQ) consiste à appliquer le principe de la modulation codée en treillis (trellis coded modulation, TCM) au codage de source. Cette technique apporte une réduction de distorsion par rapport à un quantificateur scalaire classique. De même, la TCQ permet d'améliorer les performances de la MDSQ [48].

Soit une source de mots de codes S que l'on souhaite coder en deux descriptions 1 et 2 de débits respectifs R_1 et R_2. Suivant le principe de la TCQ, les mots de codes de chacune des deux descriptions prennent leurs valeurs respectivement dans des alphabets de tailles $M_1 = 2^{R_1+1}$ et $M_2=2^{R_2+1}$. Une table d'index MDSQ de taille $M_1 \vee M_2$ permet d'associer les deux séquences S1 et S2 à la séquence S. On remarque que les mots codes de S prennent leurs valeurs dans un alphabet de taille $M_1.M_2=2^{R_1+R_2+2}$. La succession des mots codes de S1 correspond à une suite d'états dans le treillis T_1. La séquence S_1 est codée par une suite de transitions dans T_1. De même, la séquence S2 est codée par une suite de transitions dans un treillis T_2.

Par conséquent, la séquence S est codée par des paires de transitions dans les treillis T_1 et T_2. Chaque paire de transition dans ces deux treillis est identique à une transition dans le

treillis produit T =$T_1 * T_2$. Lors du codage, on cherche un chemin optimal dans le treillis T en fonction des performances souhaitées au niveau décodage du central et du décodage latéral. Plus précisément, on cherche à minimiser la distorsion centrale D sous la contrainte des distorsions latérales D_1 et D_2. La fonction à minimiser est de la forme, J = D + $\lambda_1 D_1$+$\lambda_2 D_2$.

III.7. Codage à Description Multiple basée sur codage du canal

Le codage par description multiple a été considéré comme une technique de codage conjoint source/canal [42]. Récemment, une technique de codage par descriptions multiples qui sépare le codage de source de la formation des descriptions a été proposée par Mohr et Riskin [49]. Il s'agit d'une protection inégale contre les pertes. La technique employée s'insère au niveau du codage de canal. Les sous-blocs FEC sont destinés à la correction d'erreurs. MDC a émergé comme une approche permettant d'améliorer la robustesse à des erreurs dans les systèmes de communication en général et en particulier, les réseaux internet [49].

Cette caractéristique permet d'adapter aisément le débit de transmission des données à l'état du canal ou aux possibilités du récepteur. Les flux de données ainsi crées sont dit emboités. Un tel flux est ordonné. C'est-à-dire qu'un paquet p_i ne peut être exploité que si les $i-1$ paquets $p_1 \dots p_{i-1}$ précédents ont été reçus. Il est, par conséquent, primordial de recevoir les paquets dans l'ordre. La perte d'un paquet p_i empêche le décodage de tous les paquets suivants. Les techniques de codage de canal classiques assurent une même protection sur l'ensemble des données.

La protection inégale aux erreurs a été introduite pour traiter spécifiquement les cas des flux assemblé. Ces blocs sont appliqués verticalement aux i tronçons de la i-ème couche. Grâce à ce mécanisme, chaque description contient une partie de chacune des N couches du flux initial, avec la propriété que la i-ème couche peut être récupérée à partir de i paquets quelconques.

Le codage de la source et le codage du canal étant faits séparément, il est nécessaire d'utiliser des techniques pour l'allocation de débit entre les deux types de codage [50]. La complexité de cette tâche étant très élevée, une classification est effectuée préalablement à l'allocation de la redondance, de manière à protéger en priorité les descriptions favorables. Le problème de l'allocation de redondance et de la répartition de l'information dans les descriptions et le choix du nombre optimal de descriptions est traité selon l'application.

Un des avantages de cette méthode est qu'elle peut être utilisée en conjonction avec n'importe quel codeur de source.

Les méthodes vues jusqu'à présent introduisaient de la redondance au niveau du codage de source. La méthode présentée dans [42,51] propose d'utiliser des techniques de codage de canal pour ajouter de la redondance dans le signal à transmettre. On suppose pour cela que le flux d'information à coder est organisé de façon hiérarchique et segmenté en couches d'importances décroissantes. Nous pouvons alors protéger ces couches par des codes correcteurs d'erreurs de redondance également décroissantes. Cette idée est connue sous le nom de protection inégale contre les erreurs (Unequal Error Protection, UEP).

Cette technique peut être directement décrite dans le cas général de N descriptions. L'application visée est la transmission par paquets sur un réseau où les données à transmettre sont représentées par un flux binaire hiérarchique. La configuration de paquétisation du flux est illustrée sur la figure III.6.

Les blocs sont appliqués verticalement aux i tronçons de la i-ème couche. Grâce à ce mécanisme, chaque description contient une partie de chacune des N couches du flux initial, avec la propriété que la i-ème couche peut être récupérée à partir de i paquets quelconques.

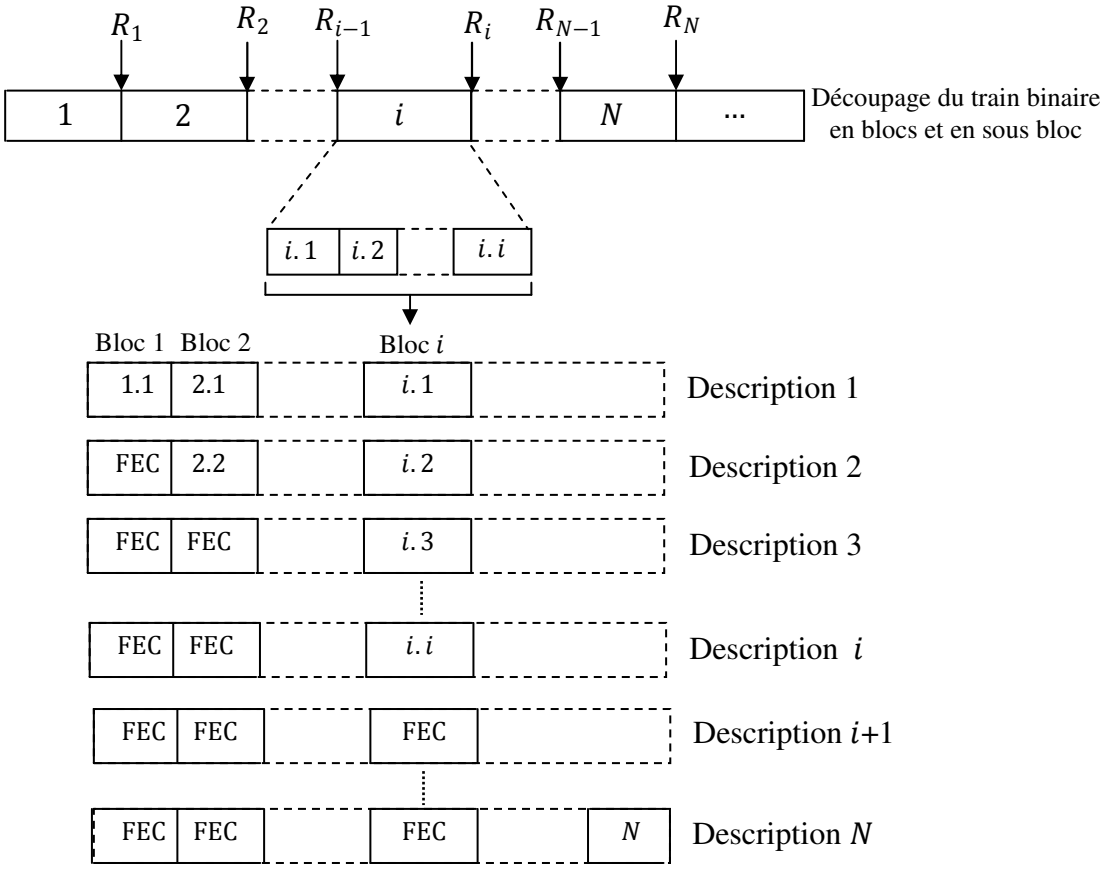


Fig. III.6. Construction de N descriptions à partir d'un train binaire

III.8. Codage par descriptions multiples basé sur des transformations (MDTC)

Dans la chaine de codage, la transformation a pour but d'une part de décorréler les données, et d'autre part d'obtenir une représentation compacte de ces données. Elle doit permettre d'assurer qu'une suppression d'une partie de l'information prouve une qualité acceptable lors de la reconstruction. Dans le cas du codage par des descriptions multiples, nous distinguons deux approches : La première consiste à remplacer la transformation classique par une autre transformation qui permet à la fois de générer les descriptions et de contrôler la redondance entre ces descriptions: La seconde approche consiste à insérer entre la transformation classique et la quantification, une autre transformation qui aura pour but de rajouter de la corrélation et de permettre la création des descriptions.

- **Codage par Descriptions multiples basées sur les trames**

Les trames peuvent aussi être utilisées pour augmenter la robustesse sur des réseaux avec pertes [44, 52]. Les trames ajoutent de la redondance dans le message à travers une expansion sur la base de fonctions redondantes (trames). Le récepteur peut alors reconstruire le message transmis sur un réseau sensible aux erreurs ou perte avec une précision suffisante si les opérateurs de la trame associée aux coefficients reçus possèdent les mêmes propriétés spécifiques. Les coefficients supplémentaires sont alors de la pure redondance et n'apportent aucune information supplémentaire.

Nous citons encore quelques transformées existantes [53]:

- Transformée multi-ondelettes.
- Codage en sous-bandes.
- Transformations par appariement de doublets.
- Transformations à base de fonctions redondantes.
- Techniques spectrales.

III.9. Réseaux et codage par descriptions multiples

Les travaux fondateurs du codage MD se basent sur un modèle « on-off » pour les canaux ou les réseaux, sans imposer de contraintes sur les délais de transmission. Plus précisément, les systèmes de codage sont conçus en considérant que chacune des descriptions est soit complètement disponible sans erreur au récepteur, soit complètement perdue. De plus, la probabilité qu'une description soit perdue est indépendante du débit choisi pour coder la source. Ces conditions sont très différentes de celles que l'on trouve en pratique sur les

canaux ou les réseaux. Par exemple, dans une situation où une communication en temps réel est requise, les descriptions pourraient être partitionnées en paquets, chacun pouvant être reçu ou perdu individuellement. De plus, la présence de congestion dépend souvent du débit de transmission. Plus le débit augmente et plus le risque de congestion augmente [54].

III.10. Les avantages du codage par description multiple

Les techniques basées sur le codage par description multiple ont pour avantage de permettre une exploitation efficace de la redondance. Elles peuvent même, comme c'est le cas des transformations basées sur les fonctions redondantes, exploiter cette redondance pour corriger une partie de l'erreur de quantification [41]:

- La séparation source/canal : permet un codage de source efficace et une mise à jour facile de ce codage de source. De plus, l'ajout de redondance est facile à quantifier, puisqu'il est indépendant de l'étape de compression. La séparation source/canal implique une conception simple et modulaire du schéma de codage. Les descriptions sont parfaitement équilibrées : chaque description à la même importance que les autres. La distorsion dépend uniquement du nombre de descriptions reçues.

- La construction optimale des descriptions: la répartition de l'information dans les descriptions peut être effectuée de manière optimale en fonction des conditions de transmission.

- Un codeur MD peut choisir de ne pas envoyer toutes les descriptions au décodeur. Par exemple, le transmetteur peut décider de ne pas envoyer certaines descriptions pour mieux exploiter la capacité du canal. C'est aussi le cas lorsque le récepteur n'a pas besoin de toutes les descriptions, ou lorsqu'il n'est pas capable de les utiliser toutes. Cette aptitude est un autre atout du codage MD.

III.11. Les inconvénients du codage par description multiple

D'une manière générale, le codage par descriptions multiples est parfaitement adapté à la prise en compte des effacements (sans mémoire) lors du codage et lors de la transmission de données sous contrainte de délai. Tant que nous ayons plus de descriptions le délai de codage augmente.

Si l'intérêt du codage par descriptions multiples a pu être montré, les comparaisons des performances avec d'autres techniques sont rares surtout dans le cas de la parole.

Un autre inconvénient réside dans la complexité du système. De plus, la mise au point de la redondance et son réglage peut s'avérer délicat. En effet, dans le cas de bancs de filtres à descriptions multiples, par exemple, il faut synthétiser un banc de filtres par niveau de redondance. Il parait difficile dans ce cas d'adapter dynamiquement la redondance en fonction des caractéristiques du canal.

Même si l'on peut imaginer des solutions permettant d'adapter la redondance, il n'en reste pas moins qu'il faut recoder complètement les données à chaque modification. En effet, les transformations se situent au début de la chaîne de codage [54]. Le réglage de la redondance s'effectue au niveau de la quantification. En revanche, lorsque toutes les descriptions sont reçues, la partie redondante n'est absolument pas exploitée.

Conclusion

Dans ce chapitre nous avons présenté la théorie du codage MD en général. Après avoir introduit la définition générale de la MDC, nous avons ensuite présenté les différentes techniques portant sur le codage à deux descriptions et sur le codage à plus de deux descriptions. Initialement, le codage par descriptions multiples a été proposé pour la transmission de données sur plusieurs canaux distincts et non corrélées. Il s'applique aussi naturellement aux transmissions sur des réseaux par paquets où les pertes de paquets sont aléatoires et indépendantes.

Les techniques de codage par description multiple basées sur les transformations ont pour avantage de permettre une exploitation efficace de la redondance surtout dans le cas de la parole. Elles peuvent même, comme c'est le cas des transformations à base de trames redondantes, exploiter cette redondance pour corriger une partie de l'erreur de quantification ou bien remplacer carrément cette trame lorsqu’elle est perdue.

Chapitre IV

Implémentation de deux codeurs MELP fonctionnant à 2.4 kbps et à 1.2 kbps

IV.1. Introduction

Nous présentons dans ce chapitre l'étude des deux codeurs LPC à excitation mixte MELP, le premier est le standard fonctionnant à 2.4 kbps, qui a été conçu par la défense américaine (DoD :Department of Defense) et le second fonctionnant à 1.2 kbps. Ces deux codeurs seront exploités pour la mise au point de notre approche MDC.

L'implémentation d'un codeur MELP comporte quatre étapes :

1. **Analyse ;** l'objet de cette composante est d'extraire les paramètres de la modélisation de la parole dans le cas du MELP.

2. **Encodeur** ; l'objectif de cette composante est double. D'abord les paramètres sont quantifiés, pour ensuite être codés et déposés dans un flux bit-stream pour la transmission.

3. **Décodeur** ; le but de cette partie est de décompresser le flux de bit transmis, corriger les erreurs de transmission détectables et de reconstruire ensuite les paramètres du modèle MELP.

4. **Synthèse** ; ce bloc permet, synthétiser le signal de la parole grâce aux paramètres du modèle décodé précédemment.

IV.2. Principe du codeur MELP

Le codeur LPC à Excitation Mixte est basé sur un modèle paramétrique, qui inclut cinq fonctionnalités améliorées comparativement aux codeurs LPC [55]. Celles-ci sont :

1. Une excitation mixte.

2. Une impulsion apériodique.

3. Amélioration spectrale adaptative.

4. Un filtre de dispersion d'impulsions.

5. Une Modélisation par les amplitudes de Fourier.

L'excitation mixte est formée de la somme d'une composante impulsionnelle et d'une composante de bruit en utilisant un mixage de filtres adaptatifs multi-bandes en vue de réduire le bruit introduit par le vocodeur LPC classique. Lorsque le signal de parole est voisé, le

codeur MELP synthétise ce signal en utilisant soit un train d'impulsions périodique ou des impulsions apériodiques. Ces dernières sont souvent utilisées dans les zones des transitions c'est-à-dire situées entre les segments voisés et les segments non-voisés du signal de parole. Ceci permet d'améliorer, lors du décodage, la reproduction des impulsions glottiques [12].

Cette excitation est une excitation multi-bande avec une intensité de voisement déterminée pour chaque bande de fréquence.

Le codeur fait une première estimation de la fréquence fondamentale, puis il calcule l'intensité de voisement dans 5 bandes de fréquence adjacentes. L'intensité de voisement est déterminée dans chaque bande par la valeur de l'autocorrection normalisée, cette dernière est codée sur 1 bit. Chaque bande est donc classée voisée ou non-voisé. L'amélioration apportée par le MELP se situe principalement au niveau de l'étage d'excitation où celle-ci devient "mixte", c'est à dire qu'elle peut prendre plusieurs formes (impulsion, bruit, régime transitoire).

Après analyse, le codeur peut positionner un indicateur appelé indicateur d'apériodicité « aperiodic flag », pour indiquer au décodeur que la composante impulsionnelle doit être apériodique ou non.

Le codeur effectue par ailleurs une analyse spectrale par prédiction linéaire et calcule les amplitudes des 10 premières fréquences harmoniques du pitch sur la transformée de Fourier du signal résiduel. Ces amplitudes sont quantifiées de manière vectorielle. L'information apportée par ces coefficients améliore la précision du filtre de production dans le domaine des basses fréquences. La qualité de la parole synthétique en présence d'un bruit de fond s'en trouve améliorée, notamment pour les locuteurs masculins [12].

Le synthétiseur interpole linéairement les différents paramètres de manière synchrone au pitch. La composante impulsionnelle est obtenue sur une période de pitch par transformée de Fourier inverse sur les 10 amplitudes de Fourier. Pour les sons non-voisés ou lorsque l'indicateur d'apériodicité est positionné, une perturbation aléatoire (jitter) est appliquée à la valeur de la période fondamentale.

Cette possibilité d'excitation impulsionnelle non périodique est particulièrement intéressante pour les zones de transition entre sons. La composante impulsionnelle et la composante de bruit sont filtrées puis ajoutées. Le filtrage appliqué à la composante impulsionnelle a pour réponse impulsionnelle la somme de toutes les réponses impulsionnelles des filtres passe-bandes pour les bandes voisées. Le filtrage de la composante de bruit est effectué de la même façon à partir des bandes non voisées. L'excitation globale est ensuite filtrée par un filtre adaptatif basé sur les pôles du filtre LPC. Son application permet le renforcement de la

structure formantique du signal synthétique.

Le signal synthétique résultant est d'abord mis à l'échelle en fonction de l'énergie de la trame originale pour ensuite être filtré dans le but est d'étaler l'énergie des impulsions sur une période de fondamentale, améliorant ainsi la qualité du signal reconstitué.

IV.3. Encodeur MELP

L'entrée de l'encodeur est un signal de parole et sa sortie est un flux de bits à transmettre. Le signal de parole d'entrée est échantillonné à une fréquence d'échantillonnage de 8 kHz et la durée de la trame est de 22.5ms correspondant à 180 échantillons. Le schéma synoptique de base et le schéma détaillé d'un codeur MELP sont donnés respectivement par la figure IV.1 et la figure IV.2. [55].

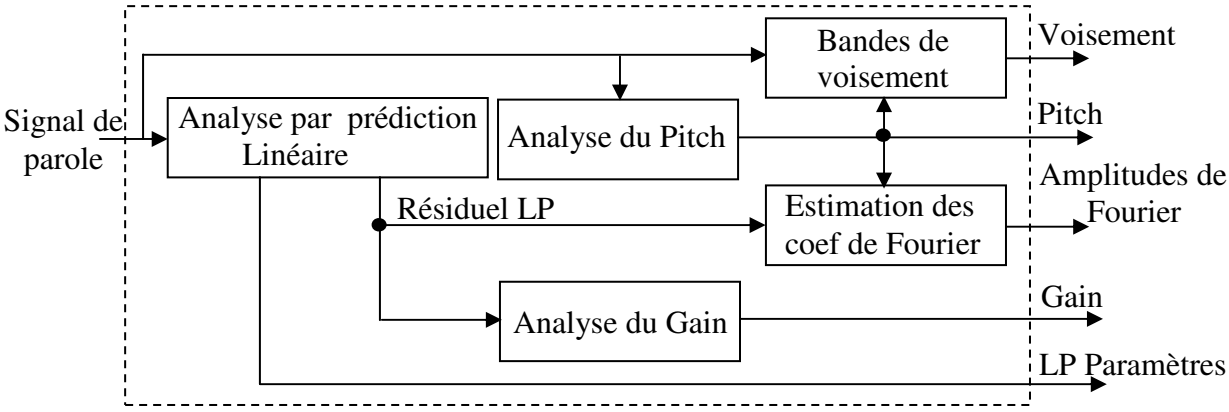


Fig. IV.1. Schéma de base du codeur MELP

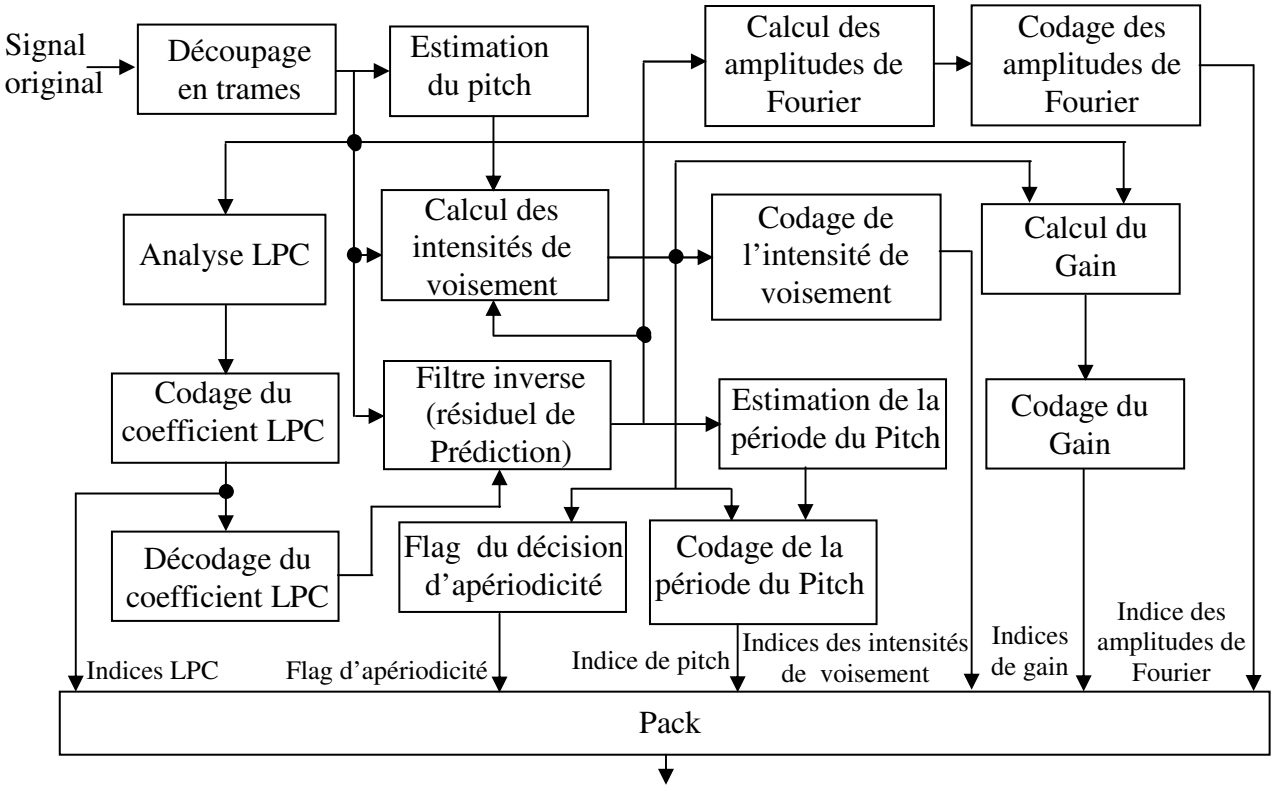


Fig. IV.2. Schéma bloc du codeur MELP

IV.3.1. Suppression des basses fréquences et calcul du pitch

La première étape dans le procédé de codage est d'éliminer toute énergie relative aux très basses fréquences qui peut être présente dans le signal d'entrée. Ceci est accompli avec un filtre passe-haut de Tchebychev d'ordre 4, ayant une fréquence de coupure de 60 Hz et un taux de rejection de 30 dB. La sortie de ce filtre est désignée sous le nom de signal de parole d'entrée dans tout ce qui suit. Avant de commencer l'estimation de la période du pitch, les positions de quelques fenêtres de traitement des signaux sont d'abord décrites. Les positions des fenêtres d'analyse sont choisies pour faciliter l'interpolation, puisque les paramètres d'une trame donnée sont interpolés entre deux ensembles différents, calculés à partir des fenêtres d'analyse [55]. Le schéma de la figure IV.3 récapitule les fenêtres principales utilisées par le codeur MELP.

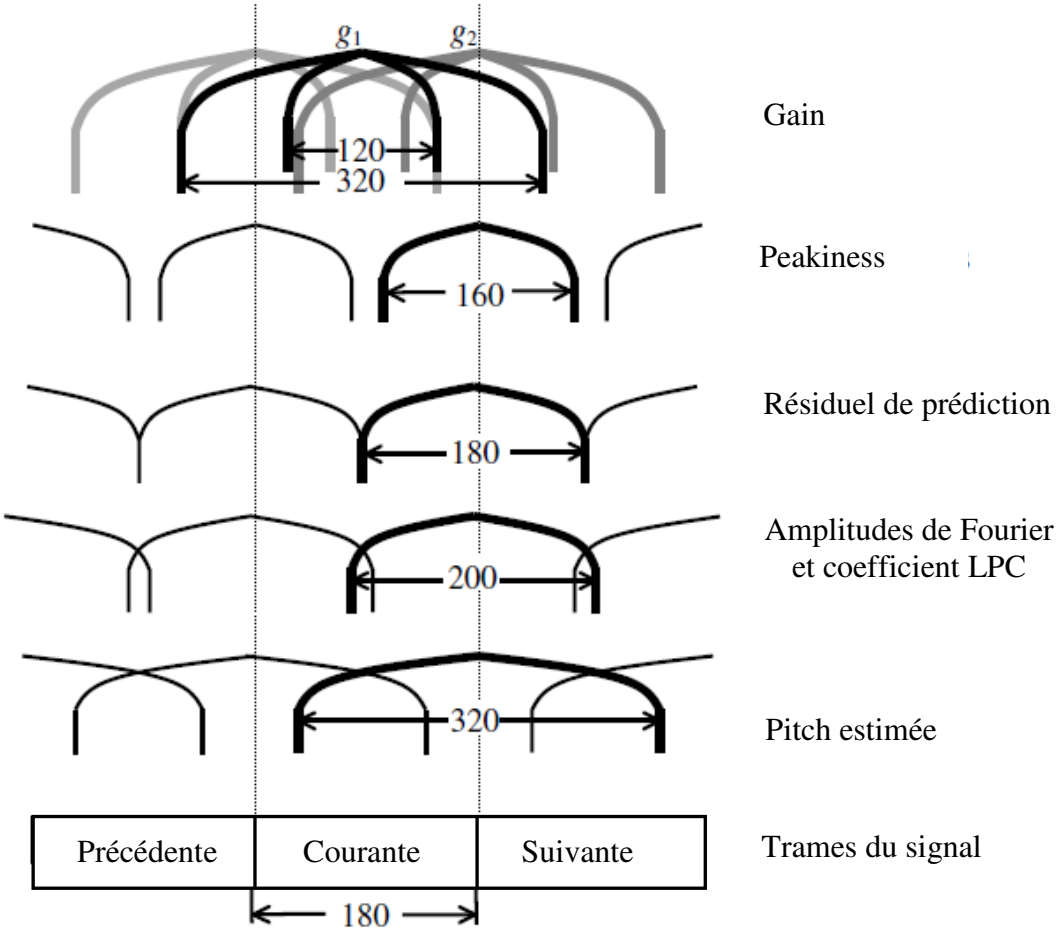


Fig. IV.3. Position des différentes fenêtres pour le codeur MELP

Pour le calcul du pitch, le signal de parole d'entrée est d'abord traité avec un filtre passe-bas de Butterworth d'ordre 6 de fréquence de coupure 1 kHz. La valeur entière du pitch, P_1, est la valeur de τ = 40, 41,…, 160, pour laquelle la fonction d'autocorrélation normalisée, r (τ), est maximale. Cette fonction est définie par [56]:

$$r(\tau) = \frac{C_\tau(0,\tau)}{\sqrt{C_\tau(0,0)C_\tau(\tau,\tau)}} \quad (4.1)$$

Avec :

$$C_\tau(m,n) = \sum_{k=-\lceil \tau/2 \rceil - 80}^{-\lceil \tau/2 \rceil + 79} s_{k+m}s_{k+n} \quad (4.2)$$

IV.3.2. Analyse de voisement et affinement du pitch fractionnaire

Cette partie de l'encodeur détermine l'intensité de voisement des cinq filtres passe-bandes, $Vbpi$, i = 1, 2,…, 5. Ce bloc permet également d'affiner la mesure du nombre entier pitch ainsi que la valeur normalisée correspondante de l'autocorrélation. L'analyse de voisement dans la bande passante commence par un filtrage de signal de parole par banc cinq bancs de filtres. Ces bancs filtres sont de type Butterworth d'ordre 6, avec des bandes passantes de [0-500], [500-1000], [1000-2000], [2000-3000] et [3000- 4000] Hz [57,58]. La figure IV.4 représente ces filtres.

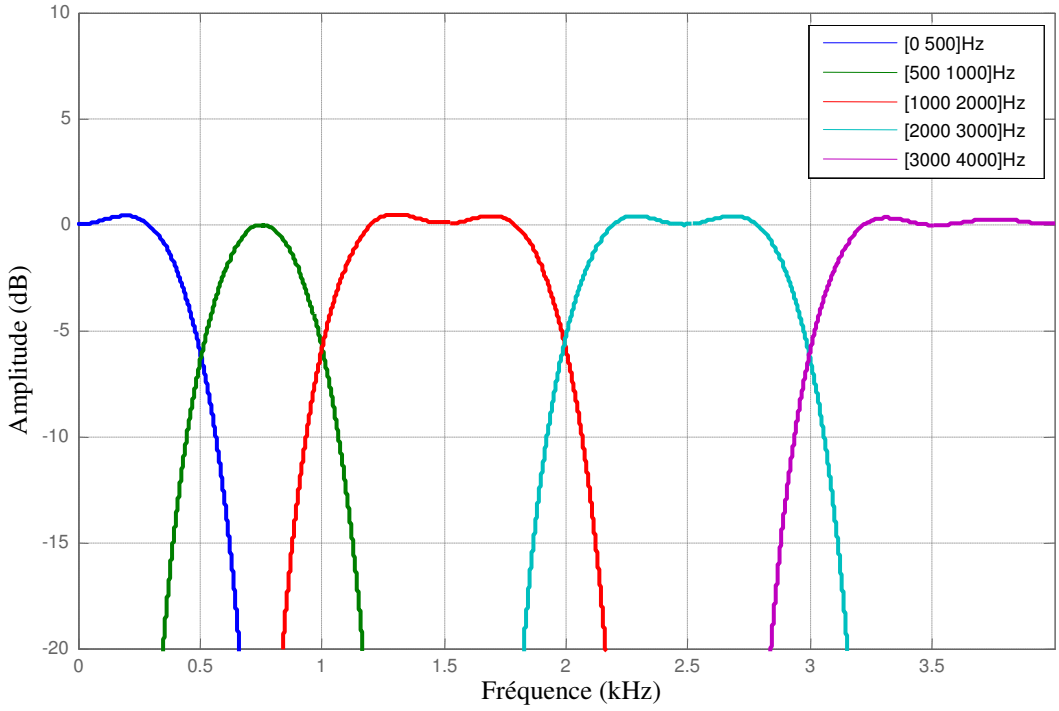


Fig.VI.4. Filtre passe bande pour l'analyse de voisement

Un raffinement de la mesure du pitch est effectué en utilisant le signal de sortie du filtre de 0-500 Hz. Cette mesure est centrée sur la sortie du filtre lorsque son entrée est le dernier échantillon dans la trame courante. Deux candidats pour le pitch sont considérés dans cette amélioration, à savoir les valeurs entières du pitch P_1 des trames courante et précédente.

Pour chaque candidat, l'équation (4.1) est utilisée pour effectuer une recherche d'un nombre entier de pitch. Un raffinement de pitch est exécuté autour du retard optimal. Ceci produit deux candidats avec leurs valeurs correspondantes prises dans l'autocorrelation normalisée. Le candidat ayant une valeur plus élevée de l'autocorrelation normalisée est choisi comme pitch fractionnaire, que l'on note P_2. L'autocorrelation normalisée correspondante $r(P_2)$ est sauvegardée comme l'intensité de voisement la plus basse dans la bande Vbp_1. La valeur de P_2 est sauvegardée pour être utilisée dans la détermination de l'intensité de voisement pour les bandes de fréquence restantes. Elle est également utilisée dans le calcul final du pitch et le calcul du gain. L'affinement du pitch fractionnaire utilise une formule d'interpolation pour améliorer la valeur du pitch de l'entrée. Cette valeur est d'abord arrondie à un nombre entier le plus proche. Ce nombre entier est supposé ayant une valeur de T échantillons. La formule d'interpolation suppose que r(τ) a un maximum entre les retards T et $T+1$. Par conséquent, $C_T(0,T-1)$ et $C_T(0,T+1)$ sont calculés et comparés pour déterminer si le maximum se trouve vraisemblablement entre T et $T+1$ ou entre $T-1$ et T. Si $C_T(0,T-1) > C_T(0,T+1)$ alors le maximum est probablement entre $T-1$ et T et le pitch T est décrémenté de 1 avant l'interpolation. L'offset fractionnaire Δ, est alors calculé par l'équation d'interpolation :

$$\Delta = \frac{C_T(0,T+1)C_T(T,T) - C_T(0,T)C_T(T,T+1)}{C_T(0,T+1)[C_T(T,T) - C_T(T,T+1)] + C_T(0,T)[C_T(T+1,T+1) - C_T(T,T+1)]} \quad (4.3)$$

Où $C_T(m,n)$ est défini par l'équation (4.2). Dans certains cas, cette formule produit un décalage en dehors de l'intervalle 0.0 à 1.0. Ainsi le décalage est maintenu entre -1 et 2. Le pitch fractionnaire est T+Δ et il est maintenu entre 20 et 160.

L'autocorrelation normalisée correspondant à la valeur fractionnaire de pitch est donnée par :

$$r(T+\Delta) = \frac{(1-\Delta)C_T(0,T) + \Delta C_T(0,T+1)}{\sqrt{C_T(0,T)[(1-\Delta)^2 C_T(T,T) + 2\Delta(1-\Delta)C_T(T,T+1) + \Delta C_T(T+1,T+1)]}} \quad (4.4)$$

La procédure de l'affinage du pitch fractionnaire est basée sur le travail présenté dans [57]. Les équations (4.3) et (4.4) produisent le décalage fractionnaire et l'autocorrélation normalisée correspondante qui seraient obtenus si le signal d'entrée avait été linéairement interpolé pour obtenir des valeurs entre les temps d'échantillonnage actuels.

IV.3.3. Analyse et codage par prédiction linéaire

Une analyse par prédiction linéaire d'ordre 10 est exécutée sur le signal de la parole d'entrée en utilisant une fenêtre de Hamming de 200 échantillons (25 ms) centrée sur le

dernier échantillon dans la trame courante. La procédure d'analyse de l'autocorrelation qui utilise l'algorithme récursif de Levinson-Durbin est mise en application. D'autre part, un coefficient d'expansion de largeur de bande de 0.994 (15 Hz) est appliqué aux coefficients de prédiction, a_i i = 1, 2,…, 10, où chaque coefficient est multiplié par 0.994i. Les Coefficients de la prédiction linéaire sont convertis en LSF.

IV.3.4. Indicateur d'apériodicité

L'indicateur d'apériodicité est mis à 1 si $Vbp_1 < 0.5$ et mis à 0 autrement. La valeur Vbp_1 est déterminée par l'analyse de voisement. Une fois fixé, cet indicateur indique au décodeur que la composante d'impulsions de l'excitation devrait être apériodique plutôt que périodique.

IV.3.5. Calcul du résiduel de prédiction et de son peakiness

Le signal résiduel de prédiction linéaire est calculé en filtrant le signal parole d'entrée avec le filtre de prédiction dont les coefficients ont été déterminés par l'analyse par prédiction linéaire. La fenêtre résiduelle est centrée sur le dernier échantillon dans la trame courante et elle est prise suffisamment large pour être utilisée lors du calcul final de pitch. Le "peakiness" du signal résiduel est calculé sur une fenêtre de 160 échantillons centrée sur le dernier échantillon dans la trame courante. Le "peakiness" est le rapport de la norme L2 à la norme L1 du signal résiduel r_n , sa valeur est donnée l'expression suivante:

$$Peakiness = \frac{\sqrt{\frac{1}{60}\sum_{n=-80}^{79} r_n^2}}{\frac{1}{60}\sum_{n=-80}^{79} |r_n|} \qquad (4.5)$$

Si le "peakiness" dépasse 1.34, alors l'intensité de voisement de la bande la plus basse, Vbp_1, est forcée à 1.0. Si le "peakiness" dépasse 1.6, alors l'intensité de voisement des trois bandes les plus basses, $Vbpi$, i = 1, 2, 3, sont aussi forcées à 1.0.

IV.3.6. Calcul définitif du pitch

La mesure définitive du pitch utilise un filtrage passe-bas du signal résiduel, où le filtre est un Butterworth d'ordre 6, avec une fréquence de coupure de 1 kHz. L'équation (4.1) est employée pour une recherche du nombre entier pitch P_2, arrondi à un nombre entier au plus proche. Une amélioration fractionnaire du pitch est alors exécutée autour d'un retard optimal d'un nombre entier du pitch. Ceci produit les valeurs expérimentales pour le pitch définitif, P_3, ce dernier est calculé suivant la procédure de calcul qui est illustrée dans l'organigramme de la figure VI.5.

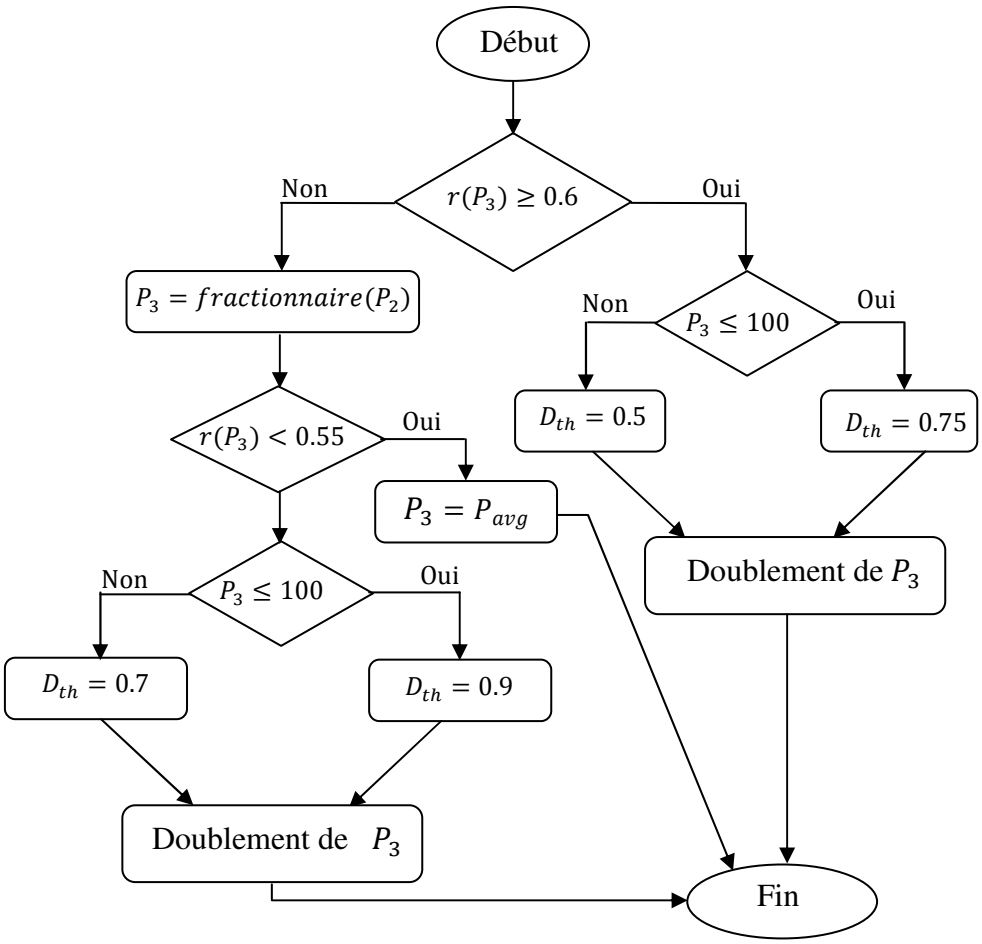


Fig.VI.5. Organigramme de calcul final du pitch

On note *r* (P_3) l'autocorrélation normalisée. En utilisant P_3 comme un pitch candidat. Le seuil de doublement D_{th} vaut 0.75 si $P_3 \leq 100$, sinon D_{th} prend la valeur 0.5. La marche à suivre de contrôle de doublement peut produire de nouvelles valeurs pour P_3 et *r* (P_3). Par contre si $r(P_3) < 0.6$, une amélioration fractionnaire du pitch autour de P_2 est exécutée en utilisant le signal de la parole d'entrée pour produire un autre P_3. Si $r(P_3) < 0.55$, alors P_3 est remplacé par P_{avg}, le pitch moyen à long terme. Sinon, la procédure de contrôle du doublement du pitch est exécutée sur le signal de parole d'entrée, en utilisant P_3 comme un pitch candidat et le seuil de doublement D_{th} prend la valeur 0.9 lorsque $P_3 \leq 100$, sinon D_{th} prend la valeur 0.7. La procédure de contrôle de doublement peut produire de nouvelles valeurs pour P_3 et *r* (P_3). Finalement, si *r* $(P3) < 0.55$, alors P_3 est remplacé par P_{avg}. Le pitch moyen à long terme, P_{avg} est mis à jour par une autre fonction du lissage simple.

IV.3.7. Contrôle de doublement du pitch

La procédure de contrôle du doublement du pitch, recherche et corrige les valeurs du pitch qui sont des multiples du pitch réel. Cette procédure considère un signal, un pitch candidat *P*,

un seuil de doublement D_{th} et renvoie le pitch contrôlé P_c et la corrélation correspondante, $r(P_c)$. Tous les calculs fractionnaires du pitch sont effectués en utilisant le signal donné à la procédure de contrôle du doublement. Cette procédure commence par une amélioration fractionnaire autour de P. Ceci produit des valeurs expérimentales pour P_c et $r(P_c)$. Ensuite, on trouve la plus grande valeur de k pour laquelle $r(P_{c/k}) > D_{th}$ $r(P_{c/k})$, où $(P_{c/k}) \geq 20$ et k= 8, 7,…, 2.

$r(P_{c/k})$ est calculée en deux étapes :

- Une amélioration fractionnaire du pitch autour de $P_{c/k}$, produisant P_k
- Une double vérification, si P_k <30. Si un tel k est trouvé, alors une amélioration fractionnaire du pitch autour de P_k est exécutée, produisant de nouvelles valeurs pour Pc et r(Pc). L'utilisation de la procédure de la double vérification assure la robustesse contre les petites valeurs fausses de pitch.

IV.3.8. Calcul du gain

Le gain est calculé deux fois par trame en utilisant une longueur de fenêtre adaptative au pitch. Cette longueur est identique pour les deux mesures de gain et est déterminée comme suit. Quand Vbp_1> 0.6, la longueur de la fenêtre est le plus petit multiple de P_2 qui est plus long que 120 échantillons. Si cette longueur dépasse 320 échantillons, elle est divisée par 2. Quand $Vbp_1 \leq 0.6$, la longueur de la fenêtre est de 120 échantillons. Le calcul du gain pour la première fenêtre produit $G1$. Il est centré 90 échantillons avant le dernier échantillon dans la trame courante. Le calcul du gain pour la deuxième fenêtre produit $G2$. Celui-ci est centré sur le dernier échantillon dans la trame courante. Le gain est la valeur de RMS, mesurée en dB, du signal S_n dans la fenêtre [59]:

$$G_i = 10\log\left(0.01 + \frac{1}{L}\sum_{n=1}^{L} s_n^2\right) \qquad (4.6)$$

Où L est la longueur de la fenêtre. Le terme 0.01 empêche l'argument du logarithme naturel de tendre rapidement vers zéro.

IV.3.9. Calcul des amplitudes de Fourier

Nous avons besoin de dix amplitudes de Fourier pour générer l'excitation périodique. Comme ces amplitudes sont celles correspondant au pitch et ses harmoniques dans le spectre de l'erreur de prédiction, la procédure pour les calculer est la suivante:

- Appliquer une FFT de 512 points sur 200 échantillons du signal de l'erreur observé par une fenêtre de Hamming, pour avoir le spectre.
- Déterminer les dix premiers pics dans le spectre du résiduel. La recherche doit être effectuée autour de pitch et ses neufs premiers harmoniques.
- Normaliser ces amplitudes en utilisant la norme L2 du vecteur qui les regroupe.

Notons que lorsque la période du pitch est faible, donc la fréquence est grande, il est possible qu'on ne puisse pas avoir dix harmoniques dans la bande de Nyquist. Dans ce cas, on retient les amplitudes qui sont dans la bande et on les complète à dix avec des 1 [60].

IV.4. Quantification des paramètres des codeurs MELP à 2.4 et 1.2 kbps

La différence entre un MELP à 2.4 kbps et un MELP 1.2 kbps se distingue au niveau de la quantification. Pour le codeur MELP à 2.4 kbps la quantification s'effectue pour des trames de 22.5 ms. Le codeur MELP à 1.2 kbps exploite les mêmes paramètres calculé par le codeur MELP à 2.4 kbps. Cependant, il regroupe trois trames consécutives pour former une super-trame de 67.5 ms, pour les quantifier globalement, afin de réduire le débit en exploitant la redondance des trames.

IV.4.1. Quantification des coefficients LSF

Dans le contexte de la compression de la parole, les coefficients de prédiction sont peu appropriés à la quantification à cause de leur large gamme dynamique et aux possibilités d'instabilité du filtre d'analyse LPC. Les coefficients LSF (Line Spectral Frequencies) ont été proposés pour palier à ces problèmes. L'objectif primordial de la quantification des paramètres LSF est la minimisation du nombre de bits attribués à ces paramètres LSF lors de la transmission. D'abord, les coefficients de la prédiction linéaire, a_i i = 1, 2,…, 10, sont convertis en fréquences de raies spectrale LSF [61]. Ensuite, un processus qui force les composantes de LSF à être dans l'ordre croissant avec une séparation minimum de 50 Hz est exécuté. Ce processus commence en contrôlant toutes les paires adjacentes des composants LSF en effectuant des permutations lorsqu'une paire quelconque ne respecte pas l'ordre croissant. Cette étape est répétée jusqu'à dix fois si cela est nécessaire. Le critère minimum de séparation est alors appliqué en corrigeant chaque paire f_i et f_{i+1} pour lesquels d =$f_{i+1} - f_i$ est moins de 50 Hz.

Le vecteur résultant LSF, f est alors quantifié en utilisant un quantificateur vectoriel à multi étages (MSVQ).

La recherche MSVQ trouve le vecteur du codebook qui réduit au minimum la distance quadratique euclidienne, d^2, entre les vecteurs quantifiés et les vecteurs de LSF non

quantifiés.

$$d^2(f,\hat{f}) = \sum_{i=1}^{10} w_i (f,\hat{f})^2 \qquad (4.7)$$

Avec la pondération suivante :

$$w_i = \begin{cases} p(f_i)^{0.3}, 1 \le i \le 8 \\ 0.64\,p(f_i)^{0.3}, i = 9 \\ 0.16\,p(f_i)^{0.3}, i = 10 \end{cases} \qquad (4.8)$$

Où $p(f_i) = \text{abs}\,(\exp(j * f_i) * LPC_i)$ et LPC_i les coefficients LPC associés.

- **Cas du codeur MELP à 2.4 kbps**

Le codebook MSVQ se compose de quatre étages de 128, 64, 64 et 64 niveaux respectivement. Le vecteur quantifie, $\hat{f}$ est la somme des vecteurs choisis par le processus de rechercher, où chaque vecteur est choisi à chaque étage.

- **Cas du codeur MELP à 1.2 kbps**

Dans le cas du codeur MELP 1.2 kbps, la quantification se fait pour chaque super-trame et tient compte du mode de voisement de chaque super-trame, en effet :

1. Si la super-trame contient une trame au plus voisée :

   Nous quantifions alors les trames séparément, chacune avec une QV simple et en utilisant un dictionnaire de 9 bits lorsque la trame est non voisée, sinon nous utilisons une MSVQ comme dans le MELP 2.4 kbps et nous utilisons le même dictionnaire de 4 étages.

2. Si la super-trame contient plus d'une trame voisée :

   Dans ce cas nous quantifions seulement la troisième trame. Les deux autres seront déduites par interpolation entre la dernière trame de la super trame précédente et celle de la super trame courante. Notons que pour la trame à quantifier, nous utilisons une MSQV de 25 bits (7, 6, 6, 6) lorsque la trame est voisée, sinon, on utilise une QV à 9 bits comme décrit précédemment (dans le premier cas). Le tableau de quantification des LSF se trouve dans le tableau. IV.1.

Tableau IV.1. Allocation des bits pour la quantification des LSF pour le codeur MELP à 1.2 kbps

Mode de voisement	LSF l_1	LSF l_2	LSF l_3	Coefficients d'interpolation	Résiduels	Totale
UUU	9	9	9	0	0	27
VUU	7.6.6.6	9	9	0	0	43
UVU	9	7.6.6.6	9	0	0	43
UUV	9	9	7.6.6.6	0	0	43
UVV	0	0	7.6.6.6	4	8.6	43
VUV						
VVV						
VVU	0	0	9	4	8.6.6.6	39

IV.4.2. Quantification du pitch

- **Cas du codeur MELP à 2.4 kbps**

La valeur finale du pitch est quantifiée sur une échelle logarithmique avec un quantificateur uniforme de 99 niveaux s'étendant de 20 à 160 échantillons. Ces valeurs du pitch sont alors élaborées dans un mot-code de 7 bits.

- **Cas du codeur MELP à 1.2 kbps**

1. Regrouper les trois pitch correspondant à une super trame dans un seul vecteur et mettre les pitch des trames non voisées à zéro.
2. Pour les super trames avec au plus une trame voisée, on quantifie ces pitch séparément en utilisant une quantification scalaire (QS) uniforme à 99 niveaux soit à 7 bits.
3. Pour les super-trames avec au moins deux trame voisées, on utilise une quantification vectorielle (QV) multi-dictionnaires pour quantifier le vecteur des pitch.

Au total, nous avons utilisé 12 bits pour la quantification du pitch et des décisions de voisement des trames. Sur les 12 bits retenus, nous avons utilisé 3 bits pour quantifier le mode de voisement (représentant les 8 cas possibles). Les 9 bits restants sont utilisés pour quantifier les valeurs du pitch. Les détails se trouvent dans le tableau. IV.2.

Tableau IV.2.Allocation des bits pour la quantification du pitch pour le codeur MELP à 1.2 kbps

Mode de voisement	3-bit	9-bit
UUU	000	• QS uniforme de 99 niveaux (7 bits) pour la trame voisée • 2 bits pour les deux autres trames.
UUV		
UVU		
VUU		
VVU	001	QV avec le même dictionnaire de dimension 512.
VUV	010	
UVV	100	
VVV	011	QV dictionnaire A 512 niveaux
	101	QV dictionnaire B 512 niveaux
	110	QV dictionnaire C 512 niveaux
	111	QV dictionnaire D 512 niveaux

IV.4.3. Quantification du gain

- **Cas du codeur MELP à 2.4 kbps**

Les deux paramètres du gain $G1$ et $G2$ sont transmis pour chaque trame. $G2$ est quantifié sur 5 bits par un quantificateur uniforme à 32 niveaux s'étendant de 10.0 à 77.0 dB. L'indice du quantificateur est transmis au mot-code. $G1$ est quantifié sur 3 bits par un algorithme adaptatif. Cet algorithme détermine si la trame est une trame équilibrée ou une trame de transition. Le tout-zéro du mot-code est envoyé pour les trames équilibrées et un quantificateur uniforme de 7 bits est utilisé pour des trames de transition. Dans ce cas-ci, l'index du quantificateur plus 1 est le mot-code transmis.

- **Cas du codeur MELP à 1.2 kbps**

Dans le codeur MELP 1.2 kbps la quantification est vectorielle avec un dictionnaire de dimension 6 et taille 1024.

IV.4.4. Quantification du voisement

- **Cas du codeur MELP à 2.4 kbps**

Quand $Vbp_1 \leq 0.6$ (non-voisé), le restant des intensités de voisement, $Vbp_i, i = 2,3,4,5$, sont quantifiés à 0. Quand $Vbp_1 > 0.6$ (voisé), le restant des intensités de voisement sont quantifiés à 1 si leur valeur dépasse 0.6 pour chacune, sinon elles seront quantifiés à 0. Cependant, il y a une exception. En effet, si les valeurs de quantification de $Vbp_i,\ i = 2,3,4,5$, 5 sont 0,0, 0, 0, 1, respectivement, alors Vbp_5 sera quantifiée à 0.

- **Cas du codeur MELP à 1.2 kbps**

Nous avons quantifié les décisions de voisement de chaque trame séparément en utilisant un dictionnaire de dimension 4. Ce dictionnaire ne contient que les combinaisons les plus probables des intensités de voisement.

IV.4.5. Quantification des amplitudes de Fourier

Pour ces amplitudes de Fourier, on utilise une pondération avec des poids variables qui favorisent les basses fréquences par rapport aux hautes fréquences avant de procéder à la quantification.

Les poids sont donnés par la formule suivante [58, 59] :

$$w_i = \left[\frac{117}{25+75\left(1+1.4\left(\frac{f_i}{1000}\right)^2\right)^{0.69}} \right]^2 \qquad i = 1,2,3,\ldots.10 \tag{4.9}$$

Où $f_i = 8000i/60$ est la fréquence en Hz correspondant au $i^{iéme}$ harmonique pour une période du pitch par défaut de 60 échantillons. Les poids sont appliqués à la différence carrée entre les amplitudes de Fourier de l'entrée et les valeurs du code-book.

- **Cas du codeur MELP à 2.4 kbps**

Les dix amplitudes de Fourier sont codées avec un quantificateur vectoriel sur 8 bits. La recherche dans le code-book est effectuée en utilisant la distance euclidienne pondérée.

- **Cas du codeur MELP à 1.2kbps**

Les amplitudes de la dernière trame voisée dans la super-trame courante sont quantifiées de la même façon que dans le MELP 2.4 kbps (QV avec un dictionnaire de 8 bits). Les amplitudes de Fourier des autres trames voisées, seront reconstituées par interpolation à l'aide du vecteur quantifié de la super-trame courante et celui de la super-trame précédente.

IV.4.6. Allocation des bits

- **Cas du codeur MELP à 2.4kbps**

Rappelons que les paramètres transmis par le codeur MELP pour reconstituer la parole synthétique sont: la fréquence fondamentale (pitch), le flag (drapeau d'apériodicité), les cinq intensités de voisement, les deux gains (correspondant aux énergies de demi trames), les dix

coefficients LPC transformés en LSF et les dix amplitudes de Fourier du pitch codées par une quantification vectorielle.

- **Cas du codeur MELP à 1.2kbps**

Pour le codeur MELP à 1.2 kbps qui fonctionnant en super-trames, on regroupe les paramètres de trois trames consécutives du MELP à 2.4 kbps. La quantification du MELP 1.2 kbps est conçue pour exploiter trois trames successives d'une part, par la quantification vectorielle (QV) pour tous les paramètres et d'autre part, la QV et l'interpolation pour les LSF, en tenant compte des propriétés de voisement et de non voisement des trames. Chaque super-trame est classée dans un codage de plusieurs états en fonction de la décision « voisement/non voisement » (V/NV) des ces trames. Le tableau IV.3 précise l'allocation de bits de tous les paramètres, les 54 bits pour une trame codée à 2.4 kbps et les 81 pour trois trames successives codées à 1.2 kbps [62].

Tableau IV.3. Table d'allocation des bits des codeurs MELP de 2.4 kbps et 1.2kbps

Paramètres	**MELP à 2.4 kbps**		**MELP à 1.2 kbps**				
Fréquence d'échantillonnage Taille de la trame Débit en trame	8kHz 180 échantillons (22.5 ms) 44,44 trames/seconde		8kHz 3*180 échantillons (67.5 ms) 14.8148 trames/seconde				
Mode de voisement	V	N/V	VVV	UVV VUV	VVU	UUV UVU VUU	UUU
10 LSFs	25	25	43	43	39	43	27
Pitch	7	7	12	12	12	12	12
10 Amplitudes de Fourier	8	-	8	8	8	8	-
5 Bandes de voisement	4	-	6	4	4	2	-
2 Gains	8	8	10	10	10	10	10
Flag	1	-	1	1	1	1	-
Protection	-	13	-	2	6	4	31
Synchronisation	1	1	1	1	1	1	1
Total de bits par trame	**54 bits**		**81bits**				
Débit total	**54*44,44= 2400 bps**		**81*14.8148 = 1200 bps**				

IV.5. Décodeur MELP

L'entrée du décodeur est un train de bits et la sortie est un signal de parole synthétisé. Le schéma synoptique de base d'un décodeur MELP est donné par la figure IV.6 et le schéma bloc détaillé est donné à la figure IV.7.

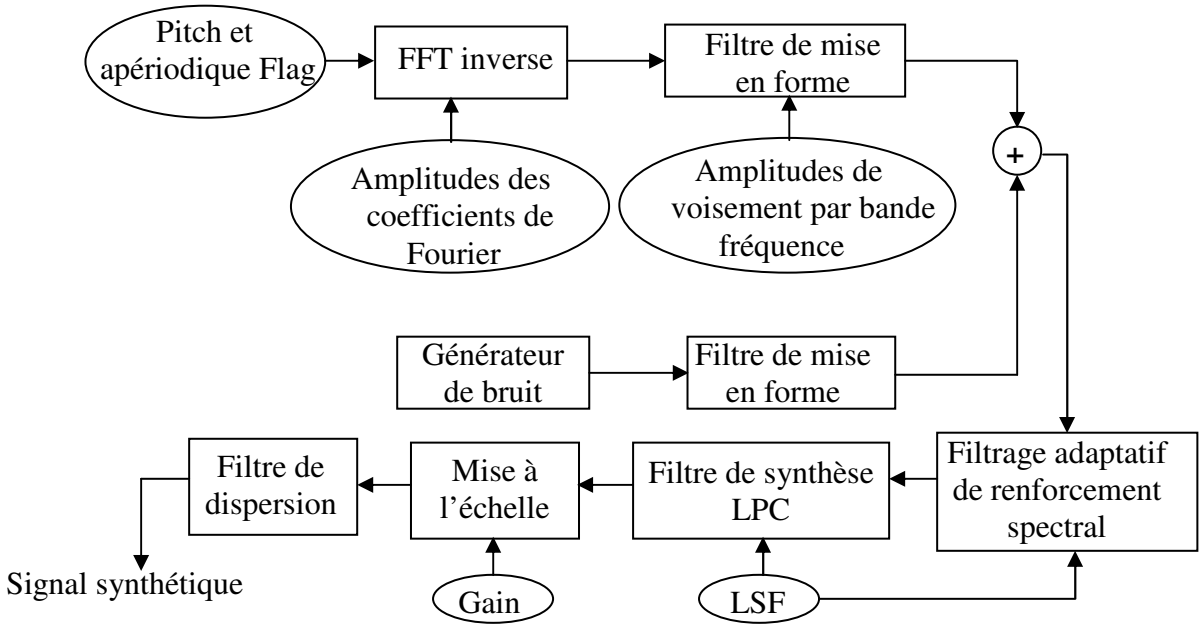


Fig. IV.6. Schéma synoptique du décodeur MELP

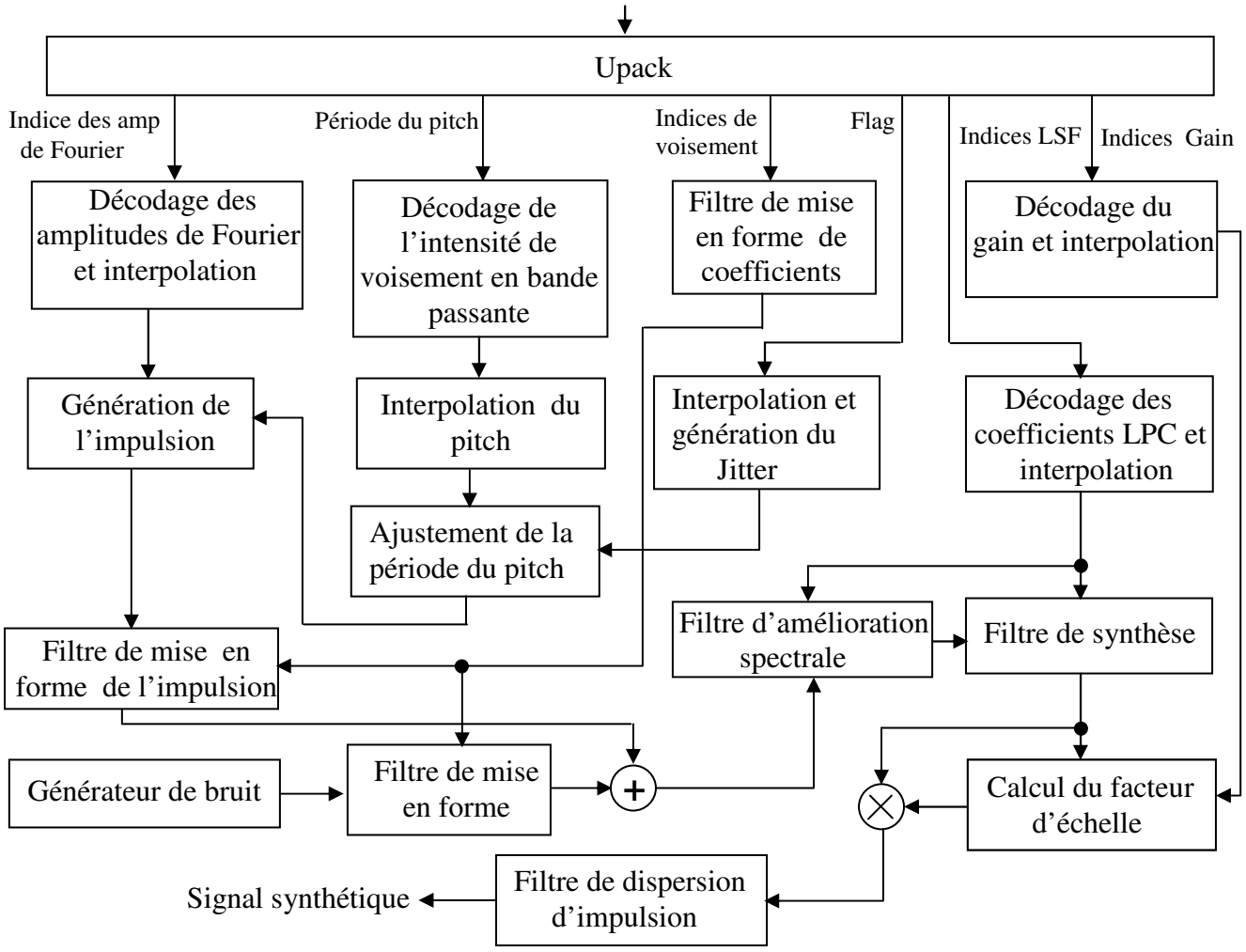


Fig. IV.7. Schéma bloc du décodeur MELP

Les bits reçus sont assemblés dans des mots-codes de chaque paramètre. Le décodage du paramètre est différent pour les modes voisés et non voisés. Le pitch est décodé en premier, puisqu'il contient l'information de mode. Les LSF sont examinés par un ordre croissant. Dans le mode non voisé, des valeurs de paramètres par défaut sont utilisées pour le pitch, la gigue (jitter), le voisement et les amplitudes de Fourier. La valeur du pitch est mise à 50 échantillons, la gigue est mise à 25%, toutes les intensités de voisement en passe-bande sont mises à 0, et les amplitudes de Fourier sont mises à 1. En mode voisé, Vbp_1 est mis à 1, la gigue est mise à 25% si l'indicateur apériodique est un 1, autrement la gigue est mise à 0%. Les intensités de voisement dans la bande passante des quatre bandes supérieures sont mises à 1 si le bit correspondant est un 1, autrement l'intensité de voisement est mise à 0.

IV.5.1. Atténuation du bruit

Pour les signaux non-voisés, une faible atténuation de gain est appliquée aux deux paramètres $G1$ et $G2$ décodés du gain qui utilise une règle de soustraction de puissance.
Avant de déterminer l'atténuation pour la première limite de gain $G1$, une évaluation de bruit de fond G_n, est mise à jour comme suit [55]:

$$\begin{cases} \text{Si } G_1 > G_n + G_{up} \text{ alors } G_n = G_n + G_{up}. \\ \text{Sinon si } G_1 < G_n - G_{down} \text{ alors } G_n = G_n - G_{down}. \\ \text{Avec } G_{up} = 0.0337435 \text{ et } G_{down} = 0.135418. \end{cases} \quad (4.10)$$

L'estimateur de bruit croit par 3dB/sec et décroit par 12dB/sec pour les taux de mise à jour du gain de 88.9 mises à jour par seconde. L'évaluation de bruit est maintenue entre 10 et 80. L'évaluation de bruit est désactivée pour les trames répétées pour empêcher l'atténuation répétée. L'évaluation de bruit de fond est également utilisée dans le calcul de l'amélioration spectrale adaptative. Le gain G_1 est modifié en soustrayant un terme (positif) de la correction, G_{att}, donné en dB par [63]:

$$G_{att} = -10\log\left(1 - 10^{0.1[G_n+3-G_1]}\right) \quad (4.11)$$

Où G_n est l'évaluation de bruit de fond (en dB), et G_1 est la première limite du gain (en dB). La correction est maintenue à des valeurs maximales de 6dB pour éviter des fluctuations au niveau du spectre et la distorsion du signal. Pour s'assurer que l'atténuation est appliquée seulement aux signaux silencieux, la valeur G_n utilisée dans l'équation (4.12) est maintenue à une limite supérieure de 20dB. Les étapes d'évaluation de bruit et de modification du gain sont alors répétées pour la deuxième limite de gain G_2.

IV.5.2. Génération d'une excitation mixte

L'excitation mixte est produite comme la somme de l'impulsion filtrée et des excitations de bruit. L'excitation impulsionnelle $e_p(n)$, n = 0, 1,…, *T-1*, est calculée par la transformée de Fourier discrète inverse sur une période du pitch [59] :

$$e_p(n) = \frac{1}{T}\sum_{k=0}^{T-1} M(k)e^{j2\pi nkT} \tag{4.12}$$

La période du pitch T est la valeur interpolée du pitch plus le temps de la gigue multipliée avec le pitch, où la gigue est la longueur de la gigue interpolée multipliée par la sortie d'un générateur de nombres aléatoires uniforme dont les valeurs sont comprises entre -1 et 1. Cette période du pitch est arrondie au nombre entier le plus proche et maintenue entre 20 et 160 échantillons. Toutes les phases pour l'excitation avec des impulsions sont mises à zéro. Par conséquent M (k) est réel. Puisque $e_p(n)$ est réel, les amplitudes obéissent à:

$$M(T-k) = M(k), \qquad k = 1,2,\dots L \tag{4.13}$$

Où L= T /2 si T est pair et L = (T-1)/2 si T est impair. M (0) est mise à 0. Les amplitudes *M (k),* k = 1, 2,…, 10, sont mises égales aux valeurs interpolées des amplitudes de Fourier et toutes les autres grandeurs qui ne sont pas spécifiées sont mises à 1. Pour empêcher des changements rapides au début de la période du pitch, l'impulsion d'excitation est circulairement décalée par dix échantillons. Ainsi, l'impulsion principale d'excitation se produit au dixième échantillon de la période. L'impulsion est d'abord multipliée par la racine carrée du pitch pour donner un signal de RMS unité et puis elle est multipliée par 1000 pour donner un niveau de signal nominal. Le bruit est produit par un générateur de nombres aléatoires uniformes avec une valeur RMS de 1000 et un intervalle s'étendant de -1732 à 1732. Les signaux d'excitation d'impulsions et de bruit sont alors filtrés et additionnés pour former l'excitation mixte. Le filtre d'impulsions de la trame courante est donné par la somme de tous les coefficients du filtre passe-bande pour les bandes de fréquence voisées, alors que le filtre de bruit est donné par la somme des coefficients de filtre passe-bande pour les bandes non voisées. Ces coefficients de filtre sont interpolés avec le pitch de façon synchrone [58]. La figure IV.8 présente le générateur le d'excitation mixte utilisé.

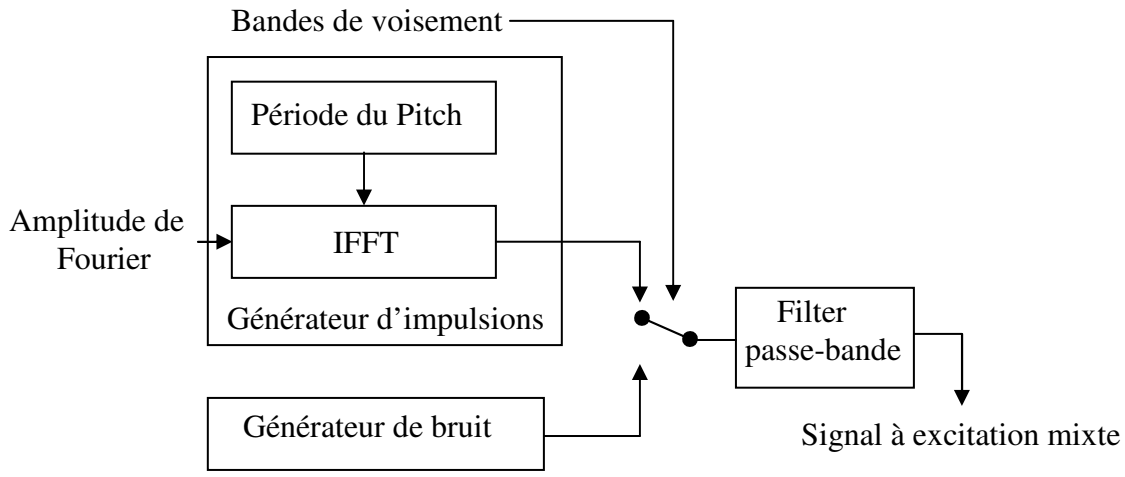


Fig. IV.8. Schéma d'un générateur d'excitation mixte

IV.5.3. Amélioration Spectrale adaptative et ajustement du gain.

Le filtre de l'amélioration spectrale adaptative est appliqué au signal d'excitation mixte. Ses coefficients sont produits par une expansion de largeur de bande de la fonction de transfert linéaire du filtre de la prédiction linéaire A(z), correspondant aux LSF interpolés. La fonction de transfert du filtre de l'amélioration $H_{ase}(z)$, est donnée par [64, 65] :

$$H_{ase}(z) = (1 - \mu z^{-1}) \frac{A(\alpha z^{-1})}{A(\beta z^{-1})} = (1 - \mu z^{-1}) \frac{1+\sum_{i=1}^{10} a_i \beta^i z^{-i}}{1+\sum_{i=1}^{10} a_i \alpha^i z^{-i}} \quad (4.14)$$

Où

$$\alpha = 0.5\rho \; et \; \beta = 0.8\rho$$

Le coefficient d'inclinaison μ est d'abord calculé en tant que max $(0.5k_1, 0)$, interpolé puis multiplié par ρ (la probabilité de signal). Par convention de signe, le coefficient de prédiction du codeur MELP, k_1, est habituellement négatif pour des spectres voisés. La probabilité ρ du signal est estimée en comparant le gain courant interpolé, G_{int} à l'évaluation du bruit de fond G_n, en utilisant la formule [55, 65] :

$$\rho = \frac{G_{int} - G_n - 12}{18} \quad (4.15)$$

Cette probabilité du signal est maintenue entre 0 et 1.

Puisque l'excitation est produite à un niveau arbitraire, le gain de la parole doit être introduit à la parole synthétisée. Le facteur de dimensionnement correct S_{gain}, est calculé pour chaque période synthétisée du pitch de longueur T.

$$S_{gain} = \frac{10^{G_{int}/20}}{\sqrt{\frac{1}{T}\sum_{n=1}^{T}\hat{S}_n^2}} \tag{4.16}$$

IV.5.4. Filtrage de dispersion

Pour éviter des discontinuités éventuelles du signal de parole synthétisée, ce facteur d'échelle est linéairement interpolé entre les valeurs précédentes et les valeurs courantes pour les dix premiers échantillons de la période du pitch. Le filtre de dispersion d'impulsions est un filtre FIR d'ordre 65, dérivé de l'impulsion triangulaire spectrale aplatie. Ce filtre a pour effet d'étaler l'énergie d'excitation d'une période du pitch. Ceci améliore la similitude entre le signal synthétique et la parole naturelle.

Le filtre de dispersion est un filtre à réponse impulsionnelle finie d'ordre 65, basé sur une impulsion glottique synthétique à spectre plat. Ses coefficients sont générés par la transformée de Fourier d'une impulsion triangulaire unitaire [66]. La figure IV.9 et la figure IV.10 illustrent sa réponse fréquentielle et sa réponse impulsionnelle.

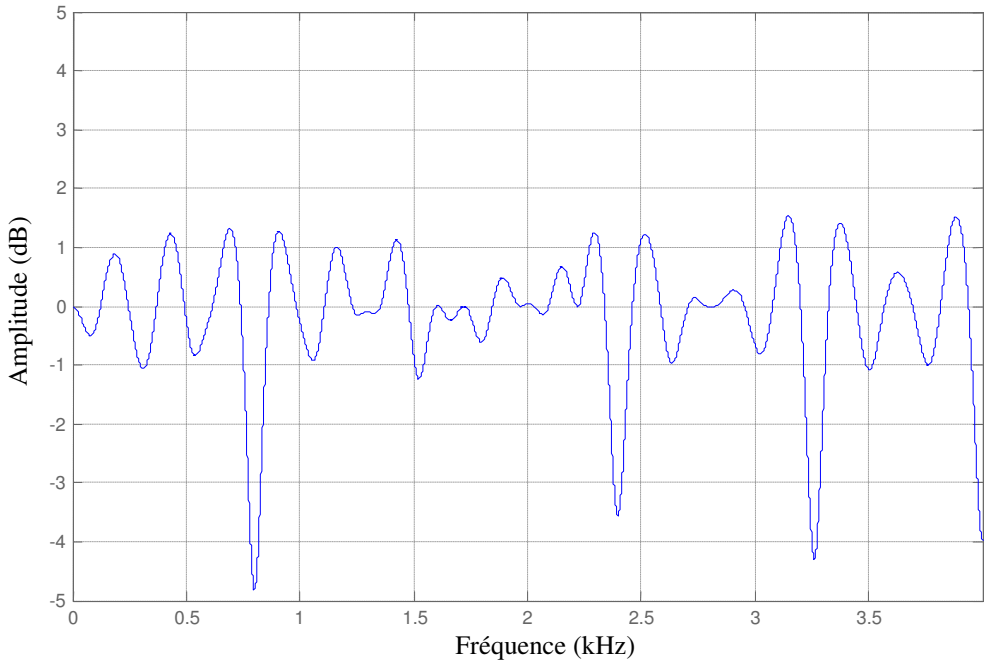


Fig. IV.9. Réponse fréquentielle du filtre de dispersion

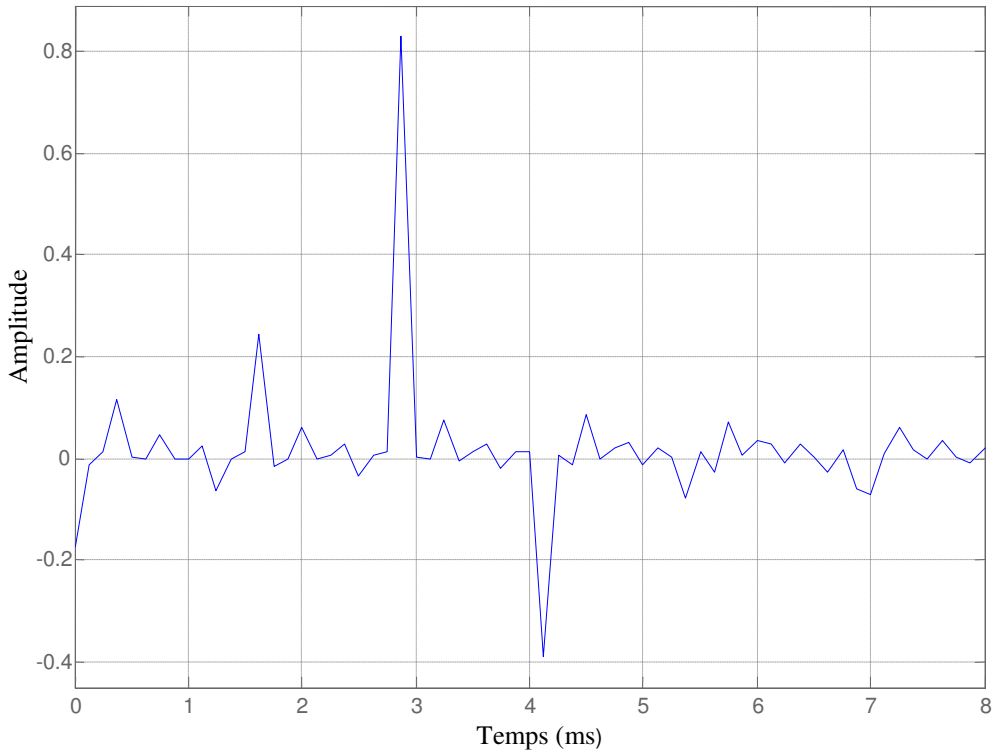


Fig. IV.10. Réponse impulsionnelle du Filtre de dispersion

Conclusion

Nous avons présenté dans ce chapitre, notre mise au point de deux codeurs MELP fonctionnant respectivement à 2.4 kbps et à 1.2 kbps en marquant les points de différence au niveau de la quantification des paramètres à coder pour chaque codeur, car le premier codeur la quantification se fait sur une trame de 22.5 ms, alors que le second codeur utilise une approche multi-trames (trois trames consécutives). Le but est alors de quantifier vectoriellement les paramètres de la troisième trame contenue dans la super-trame courante pour les utiliser dans une interpolation, afin de déduire les paramètres des deux trames qui la précèdent en profitant de la redondance existant dans le spectre du signal de parole. Ceci permet d'avoir un grand gain en débit.

Chapitre V

Résultats et Evaluations

V.1. Introduction

Nous présentons dans ce chapitre, l'implémentation de la technique de codage par description multiple mise au point utilisant les deux codeurs MELP fonctionnant respectivement à 2.4 kbps et 1.2 kbps. Le but est de combattre les pertes de paquets lors d'une transmission de la parole sur le réseau IP. Pour évaluer les résultats (la qualité perceptuelle du signal reconstruit), nous avons choisi les méthodes objectives, connues sous le nom mesure perceptuelle objective, basées sur la modélisation de l'audition et qui tient compte de certaines caractéristiques de la perception humaine. Dans un premier temps, nous avons évalué le quantificateur des coefficients LP à l'aide de la mesure objective de la distorsion spectrale et le la mesure du PESQ, afin d'évaluer les performances de ces deux codeurs MELP. Par la suite, nous avons procédé à l'évaluation de la MDC en utilisant toujours la méthode objective PESQ. Les différents tests effectués et les résultats les obtenus sont illustrés et commentés.

V.2. Evaluation de la qualité perceptuelle de la parole

La qualité vocale peut être mesurée en utilisant des méthodes objectives ou subjectives. La mesure subjective (par exemple, MOS) est la référence en matière des méthodes subjectives, mais elle est lente et difficile à mettre en œuvre. Ainsi, les notes des participants pour une condition de test donnée sont moyennées pour obtenir la note moyenne d'opinion (MOS), qui permet de diminuer l'effet subjectif sur l'évaluation de la qualité vocale. De plus, la perception de la qualité vocale dépend du contexte et de l'environnement dans lesquels est placée la personne qui juge. De même, l'environnement (bruit, informations visuelles ou sonores supplémentaires) influence le jugement de la qualité. Ainsi, les conditions à tester sont définies en fonction de l'objectif visé. Le participant est amené à évoluer dans un ou plusieurs contextes (écoute, locution et conversation).

Actuellement, l'évaluation de la qualité de la parole utilise une méthode dite PESQ, normalisée ITU-T, conjointement à des tests d'écoute.

V.2.1. Mesure de distorsion spectrale (SD)

La distorsion spectrale est une mesure objective, souvent utilisée pour évaluer la performance des quantificateurs dans les codeurs fonctionnant à très bas débit. Cette mesure de distorsion donne une bonne corrélation avec la perception humaine. La SD se calcule sur un intervalle de bande limité, son expression, est donnée, pour une i trame par [12] :

$$SD_i = \sqrt{\frac{10}{n_1 - n_0} \sum_{n=n_0}^{n_1} \left[\log_{10} \left(\frac{S(f_n)}{\hat{S}(f_n)} \right) \right]^2} \qquad (5.1)$$

$S(f_n)$ et $\hat{S}(f_n)$ sont respectivement les spectres de puissance originaux $S(f_n)$ et quantifiés $\hat{S}(f_n)$ du filtre de synthèse LPC. Une FFT sur 256 points est appliquée pour calculer les deux spectres de puissance $S(f_n)$ et $\hat{S}(f_n)$ de la ième trame du signal de parole.

La SD moyenne est évaluée pour toutes les trames des fichiers traités, son expression donnée par :

$$\overline{SD} = \frac{1}{M} \sum_{i=1}^{M} SD_i \quad , \quad M \text{ nombre de trames} \qquad (5.2)$$

En général, une $\overline{SD}$ d'environs 1dB indique que la distorsion perçue pendant la quantification est négligeable. Paliwal et Atal ont établi que la SD moyenne n'est pas suffisante pour mesurer transparence d'un quantificateur. Ils ont introduit la notion des trames spectrales externes ou (outliers frames). Par conséquence nous pouvons obtenir une qualité de codage transparent si les conditions suivantes sont maintenues [12].

- La SD moyenne est d'environ 1dB
- Le pourcentage des trames ayant une SD entre 2 et 4 est moins de 2%
- Aucune trame ne doit avoir une SD qui dépasse 4 dB

V.2.2. Evaluation perceptuelle de la qualité vocale (PESQ)

Pour tester la qualité de la parole obtenue par la méthode proposée, nous avons utilisé la méthode d'évaluation objective normalisée dite PESQ (perceptual evaluation of speech quality). Celle-ci a été décrite dans la recommandation P.862 de l'ITU-T (union internationale des télécommunications), [67, 68]. La méthode PESQ est utilisée pour la prédiction de la qualité subjective pour la téléphonie et pour les codeurs vocaux. Elle est destinée à évaluer l'influence de certains facteurs tels que la perte de paquets, le délai variable et les distorsions

dues aux erreurs de canal et qui sont mal évaluées par les méthodes classiques. La méthode PESQ est conçue pour comparer une version de référence (originale) à celle obtenue par synthèse à partir de cette référence, mais après transmission ou après avoir subi des dégradations. L'idée de base de PESQ est de transformer les formes d'onde en une représentation perceptuelle, similaire à la représentation des paramètres utilisés par les vocodeurs à bande étroite. En d'autres termes, la différence entre les paramètres de ces représentations est évaluée à l'aide d'un modèle cognitif en vue d'estimer la distance de perception entre les deux signaux : le signal original et le signal synthétique (décodé) (figure V.1).

Pour donner un aperçu sur la position des limites entre la qualité « bonne » et « basse » et entre celle « basse » et celle « inacceptable » la recommandation P. 862 donne les intervalles de la valeur de PESQ suivants [68]:

- PESQ donne une valeur numérique entre 0(aucune similitude) et 4.5 (signaux identiques), qui simule la perception humaine de la qualité de la parole.
- Des scores PESQ entre 3 et 4.5 désignent une qualité très acceptable (avec un 3.8 comme seuil de la qualité dans les systèmes téléphonique traditionnels), niveau qu'on va appeler qualité bonne.
- Des valeurs entre 2.5 et 3 indiquent une qualité acceptable entre 2 et 2.5, niveau de qualité est dit bas. Un effort est alors nécessaire pour la compréhension. Dans ce, cas on va référer à une qualité dite basse.
- Des valeurs inférieures à 2 signifient que la dégradation a rendu la communication très difficile ou impossible. En d'autre terme, l'intelligibilité est perdue ; par conséquent, la qualité est inacceptable.

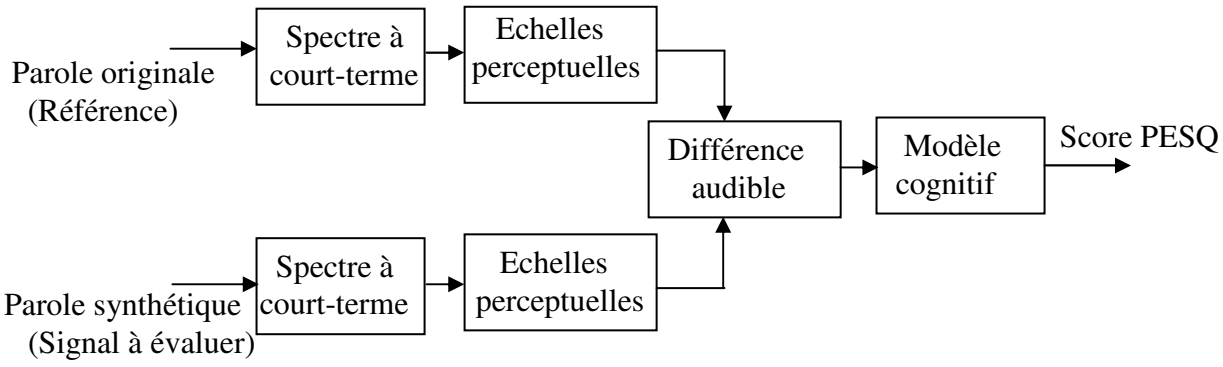


Fig. V.1. Schéma synoptique permettant l'estimation la distance perceptuelle PESQ

V.3. Description des signaux de parole utilisés dans les tests

Pour tester et valider notre méthode, nous avons utilisé un matériau linguistique formé de corpus multilingue. Le premier est composé de phrases arabes phonétiquement équilibrées [69] conçu au niveau de notre laboratoire. Ce corpus contient un total de 60 phrases, 10 phrases prononcées par 3 locuteurs féminins et 3 locuteurs masculins. La fréquence d'échantillonnage du signal parole des fichiers était de 10 kHz, nous avons dues effectuer un sous-échantillonnage de toute la base de données à 8 kHz, pour mettre dans les conditions de la téléphonie. Pour les langues française et anglaise, nous avons utilisé les phrases célèbres phonétiquement équilibrées : « la bise et le soleil » et « the sun and wind»

V.4. Evaluations des codeurs MELP implémentés

L'utilisation de codeurs pour envoyer de la voix sur un canal de communication induit, par le mécanisme de compression des données, il y a une baisse de la qualité perçue ; par conséquent pour chaque codeur il y a un score PESQ maximal qui peut être obtenu. Il est évident que sa performance quand des dégradations apparaissent dans le réseau doit seulement être estimée par rapport à ce score maximum. Nous allons évaluer les performances de ces codeurs MELP fonctionnant à 2.4 kbps et 1.2 kbps respectivement, la première évaluation consiste à la mesure de distorsion spectrale du quantificateur des LSF, par la suite, nous mesurons la qualité perceptuelle de chaque codeur en utilisant le PESQ.

V.4.1. Evaluation du quantificateur des LSF pour les deux codeurs avec SD

Afin d'évaluer la technique quantification, pour les deux quantificateurs (MSVQ pour le codeur à 2.4 kbps et la quantification vectorielle conjointe avec interpolation pour le codeur à 1.2 kbps) proposée pour quantifier les coefficients LP. Nous avons utilisé la mesure de la distorsion. Pour 1775 trames analysées nous avons obtenu les résultats suivants :

Tableau V.1. Résultats du test d'évaluation de SD

	SD moyenne (dB)	**Trames>2dB** (%)	**Trames> 4dB** (%)
MELP à 2.4 kbps	1.03	4.40	1.77
MELP à 1.2 kbps	1.61	19.6	6.57

Pour le premier codeur, les résultats sont satisfaisantes, il ya moins de distorsion. Pour le second codeur, la distorsion moyenne est supérieure à 1. Cela est dû à la double erreur de quantification vectorielle et à l'interpolation. Cela implique une perte d'information au niveau de la quantification. Cette perte est nécessaire pour réduire le débit.

Les coefficients LP influencent beaucoup la qualité naturelle, mais pas son intelligibilité. Comme dans notre cas, le MELP à 1.2 kbps est employé dans la technique MDC juste pour combler les trames perdues, donc sa contribution, dans la restitution du signal synthétique est moins significative comparativement au codeur MELP à 2.4 kbps.

V.4.2. Evaluation de la qualité par le PESQ pour les deux codeurs MELP

Nous évaluons les performances des deux codeurs MELP implémentés séparément, Le but est de chiffrer la qualité perceptuelle de nos codeurs avant qu'ils soient implémentes pour la MDC. Pour cela nous avons pris une vingtaine de dans le corpus arabe PAPE. Nous avons calculé le score PESQ pour ces phrases avec les codeurs MELP, nous avons estimé la moyenne de ces scores pour les locuteurs masculins et les locutrices féminines. Nous resumons les resultats moyen dans ce tableau.

Tableau V.2. Résultats des tests objectifs de deux codeurs MELP

	MELP à 2.4 kbps (PESQ)	MELP à 1.2 kbps (PESQ)
Locuteurs masculins	**2.99**	**2.61**
Locutrices féminines	**2.89**	**2.33**

D'après les résultats obtenus, nous constatons que :

Le score PESQ obtenu est meilleur un les locuteurs masculins que les locutrices féminines pour les deux codeurs MELP. La perte de la qualité observée pour le MELP à 1.2 kbps en comparaison avec MELP à 2.4 kbps est un résultat prévu.

V.5.Technique de recouvrement des trames perdues

Dans ce travail, nous avons mis au point une méthode pour combattre les pertes de paquets basée sur la technique MDC, c'est-à-dire incluant une redondance. L'information redondante n'a pas la qualité de l'information originale car elle est grossièrement quantifiée, mais elle contribue à reconstruire la parole lorsqu'il ya perte de paquet. Notre approche consiste en une paquétisation qui se fait au niveau des trames, où chaque paquet contient à la fois des informations sur la trame courante et des informations sur les trois trames adjacentes à venir. Ce sont ces redondances que l'on appelle descriptions [70].

En effet, nous avons codé le signal sur deux descriptions. La première utilise un codeur MELP et permet de coder la trame courante Tn sur 2.4 kbps. Elle sert à la reconstruction du signal avec une bonne qualité. La seconde utilise un autre codeur MELP, mais fonctionnant à

un débit de 1.2 kbps. Ce dernier code trois trames successives dans le même paquet, à savoir les trames **Tn+1, Tn+2 et Tn+3** qui succèdent à la trame Tn. Notons que le MELP 2.4 utilise une trame de 22.5 ms alors que le MELP 1.2 opère sur une trame de 67.5 ms, soit trois trames du MELP 2.4. (Vu dans le chapitre IV).

La multi-description ainsi constituée contribue à reconstruire la parole lorsqu'un, deux, trois, voire quatre paquets successifs seront perdus.

V.6. Format d'un paquet

Pour le format d'un paquet, nous proposons d'inclure les deux descriptions comme indiqué à la figure V.2. La première description inclut une quantification assez fine à 2.4 kbps, c'est-à-dire celle donnée par le standard MELP 2.4 kbps. Cette description est dédiée à la trame courante **Tn**. Elle est nécessaire pour procurer une bonne qualité de la parole en condition de non erreur (sans perte). La deuxième description est grossièrement quantifiée à 1.2 kbps. Elle contient trois trames successives à la trame courante : **Tn+1**, **Tn+2** et **Tn+3**. Elle procure une qualité raisonnable et sert à recouvrer jusqu'à trois pertes de paquets.

Le total de bits alloués à un paquet sera donc de 135 bits et l'allocation de ces bits est illustrée dans le tableau IV.3. On obtient un débit total de transmission de 6 kbps, correspondant à la longueur d'une trame de 22.5 ms.

Comparativement aux travaux faits sur la VoIP avec la MDC avec le codeur G. 729, nous avons un gain de 2 kbps en débit.

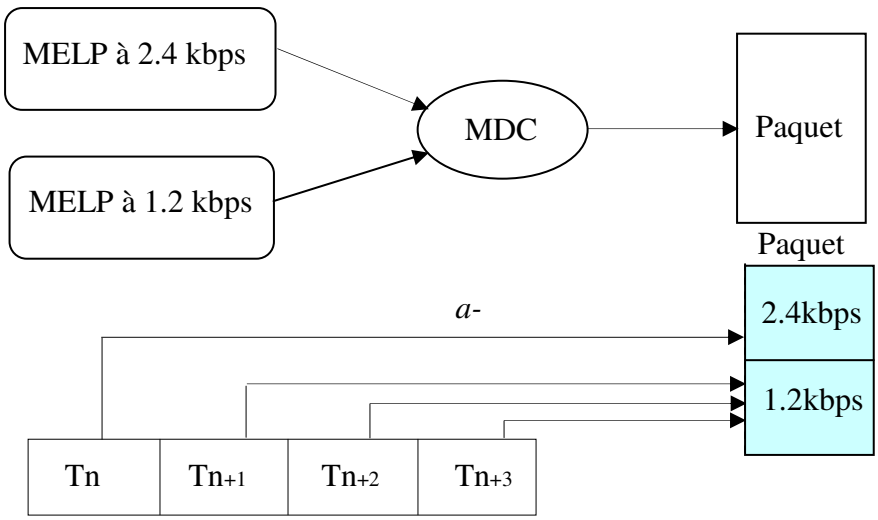


Fig. V.2. Schéma synoptique de notre paquétisation utilisant 2 descriptions:

a- Formation d'un paquet à l'aide de 2 codeurs MELP.

b- Affectation des trames

La figure V.3 montre comment le système MDC permet de recouvrer les paquets perdus. Si un ou deux ou trois de l'ensemble de paquets sont perdus, cette configuration nous permettra de récupérer le (les) paquet (s) perdu (s). Ainsi, jusqu'à 3 paquets peuvent être recouvrés. On peut même tenter de récupérer la quatrième trame, au cas où le paquet N°14 sera également perdu, en procédant par une méthode d'extrapolation. Sur cette même figure, nous avons représenté, de gauche à droite, les cas respectifs des pertes de 1 paquet, de 2 paquets, de 3 paquets et enfin de 4 paquets.

Le nombre de paquets pouvant être recouvré peut aller jusqu'à trois. Une quatrième trame peut même être recouverte en utilisant une technique d'extrapolation. On observe sur cette figure les cas suivants :

- Le 1er cas correspond à seul un paquet perdu (trame T_2). Le codeur à 2,4 kbps ne pouvant plus nous fournir la parole, on se rabat sur la trame précédente qui contenait dans son paquet les informations permettant de reconstituer trois trames T_2, T_3 et T_4. On récupère directement la trame T_2 de ce paquet.

- Le 2ème correspond à deux paquets successifs perdus, à savoir T_4 et T_5. Comme précédemment, on ne peut alors construire le signal à partir d'un MELP à 2,4 kbps. De même que précédemment, on procède à leur récupération du paquet précédent, reçu correctement.

- Dans le troisième cas, lorsque trois paquets successifs T_7, T_8 et T_9 sont perdus, au décodage du signal on procède toujours à leur récupération de la trame passée.

- Dans le dernier cas, lorsqu'on perd jusqu'à quatre paquets successifs T_{11}, T_{12}, T_{13} ces trois trames seront récupérées dans le paquet N°10. Ce paquet contient à la fois les informations sur la trame T_{10} codée à 2.4 kbps et celles concernant les trames suivantes T_{11}, T_{12} et T_{13} où l'ensemble est codé à 1.2 kbps. La trame T_{14}, sera récupérer les trois premières à partir du paquet précédent et de par la méthode d'extrapolation.

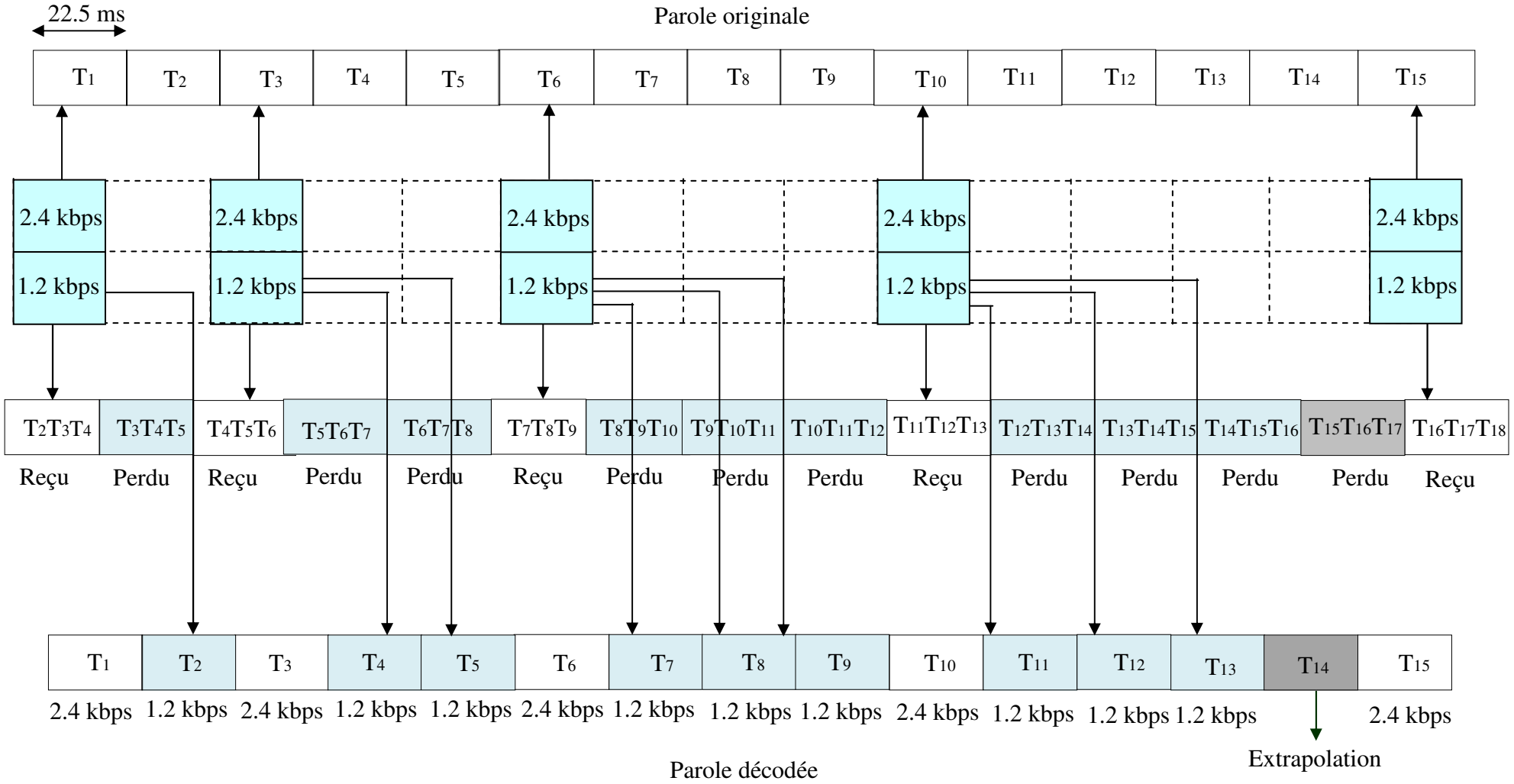


Fig. V.3. Processus de recouvrement de paquets, basé sur la MDC

V.7. Déroulement de tests

Cette section présente les simulations utilisées pour effectuer nos tests. Nous avons simulé différentes pertes pour introduire une dégradation au niveau du signal synthétique. Ces pertes ont été simulées de façon aléatoire par utilisation de la fonction RAND(X) qui suit une loi de distribution uniforme.

Le taux de perte des paquets est donné par la formule suivante :

$$\text{Taux} = \frac{\text{nombre de trames perdues}}{\text{nombre de total de trames}} \text{ x } 100 \qquad (5.3)$$

Nous calculons aussi le score PESQ par comparaison des signaux de parole de sortie avec les fichiers de parole de référence. Dans notre cas, nous donnons deux scores PESQ : le premier concerne le signal original et signal synthétique ayant subi des pertes, le second concerne entre le signal original et celui synthétique après récupération des trame perdues grâce à la MDC.

Nos tests consistent à faire neuf expériences pour chaque phrase. Pour chaque expérience nous avons calculé les valeurs obtenues lorsqu'on varie des taux de perte entre 0 et 30%. Ceci permet de bien montrer l'évolution de la qualité en fonction du taux de perte, comme mentionné dans les tableaux V.3, V.4 et V.5.

Nous avons testé la technique MDC par simulation sur ordinateur et nous avons observé qu'une redondance, même de valeur faible, contribue à accroître la robustesse vis-à-vis des pertes de paquets. Nous avons ensuite utilisé différents taux de perte de paquets et testé la robustesse du codeur MELP utilisant la MDC.

Nous avons utilisé l'environnement MATLAB pour la simulation. La figure V.4 représente le schéma général de simulation.

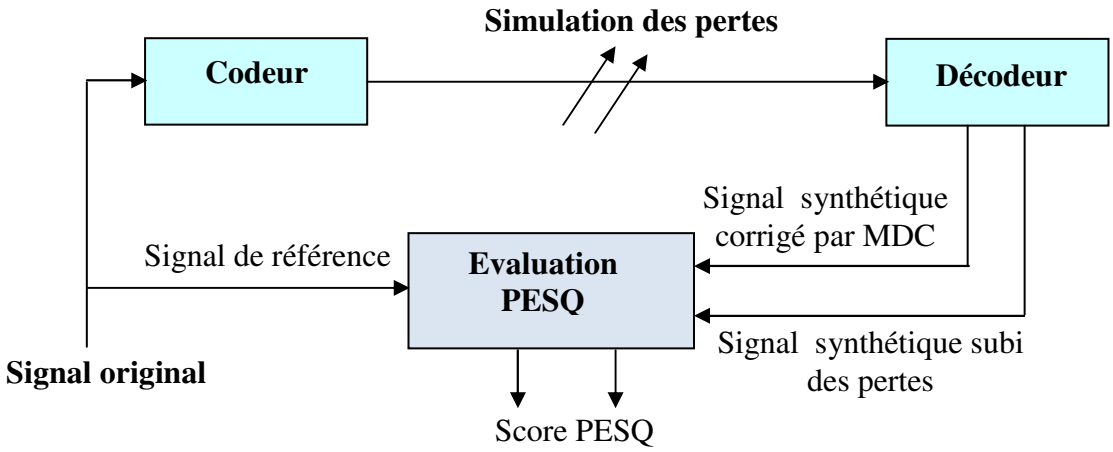


Fig. V.4. Schéma de la simulation

V.8. Evaluation des résultats de la MDC

Les tableaux et les graphes donnés dans ce paragraphe donnent les taux de perte des trames chiffrés en fonction des valeurs PESQ pour les deux cas c'est-à-dire avant et après l'application de la MDC. Nous avons pris vingt phrases prononcées par des locuteurs masculins et vingt phrases prononcées par des locutrices féminines. Ces phrases sont prises dans la base de données PAPE (Phrases Arabes Phonétiquement Equilibrées). Nous donnons à chaque fois la valeur moyenne des PESQ pour chaque valeur de taux de perte.

Dans les tableaux. V.3, V.4 et V.5 nous résumons les résultats obtenus à la suite des tests effectués sur les vingt phrases prononcées d'abord par des locuteurs masculins, ensuite par des locutrices et enfin en combinant les phrases des deux types de locuteurs. Les figures V.5, V.6 et V.7 représentent l'évolution de la qualité observés en fonction des taux de perte des paquets.

Tableau. V.3. Comparaison entre le PESQ obtenu par le MELP avant et après application de la technique MDC, pour différents taux de perte pour des locuteurs.

Taux de perte de trame en %	MELP sans MDC (PESQ)	MELP avec MDC (PESQ)	Variation du PESQ
0	2.99	2.99	0
5	2.75	2.90	0.15
10	2.59	2.80	0.21
12	2.47	2.74	0.27
15	2.33	2.68	0.35
18	2.21	2.52	0.31
20	1.96	2.44	0.48
25	1.51	2.35	0.84
30	1.38	2.26	0.88
Moyenne			**0.49**
Ecart type			**0.79**

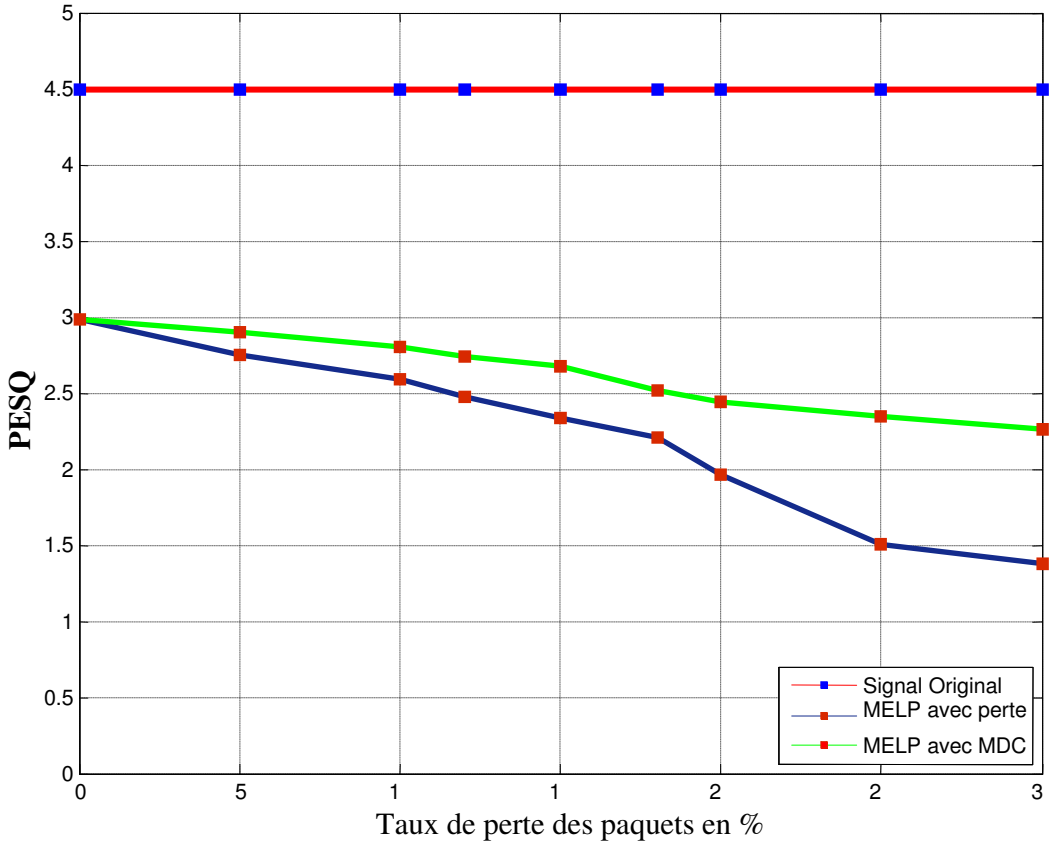


Fig. V.5. Evolution objective de la qualité en fonction des pertes par paquets dans le cas des locuteurs

Tableau .V.4. Comparaison entre le PESQ obtenu par le MELP avant et après application de la technique MDC, pour différents taux de perte dans le cas des locutrices.

Taux de perte de trame en %	MELP sans MDC (PESQ)	MELP avec MDC (PESQ)	Variation de PESQ
0	2.89	2.89	0
5	2.70	2.80	0.10
10	2.54	2.69	0.15
12	2.29	2.57	0.28
15	1.99	2.39	0.40
18	1.41	2.20	0.79
20	1.22	2.12	0.9
25	1.01	2.05	1.04
30	0.97	1.98	1.01
Moyenne			**0.58**
Ecart type			**0.39**

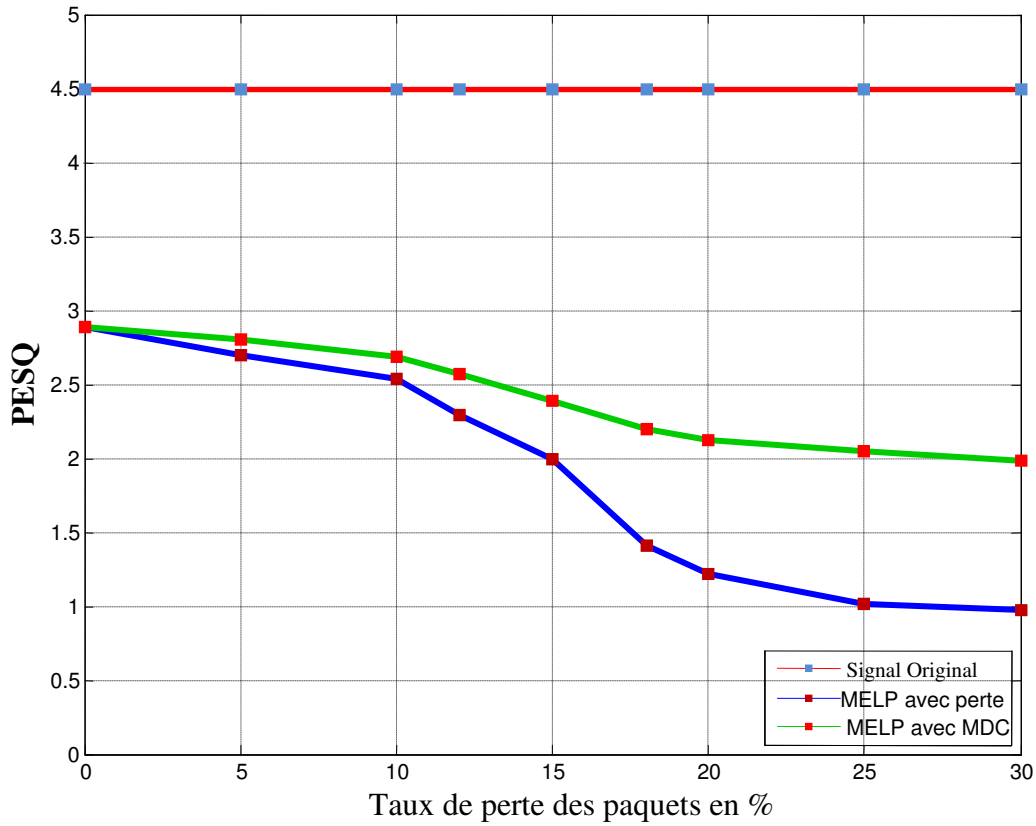


Fig. V.6. Evolution objective de la qualité en fonction des pertes par paquets dans le cas des locutrices.

Tableau .V.5. Comparaison entre le PESQ obtenu par le MELP avant et après application de la technique MDC, pour différents taux de perte pour le cas combiné des locuteurs et locutrices.

Taux de perte de trame en %	MELP sans MDC (PESQ)	MELP avec MDC (PESQ)	Variation de PESQ
0	2.92	2.92	0
5	2.73	2.85	0.12
10	2.58	2.81	0.23
12	2.43	2.74	0.31
15	2.11	2.61	0.50
18	1.69	2.45	0.69
20	1.44	2.38	0.94
25	1.25	2.30	1.05
30	1.12	2.11	0.99
Moyenne			**0.60**
Ecart type			**0.33**

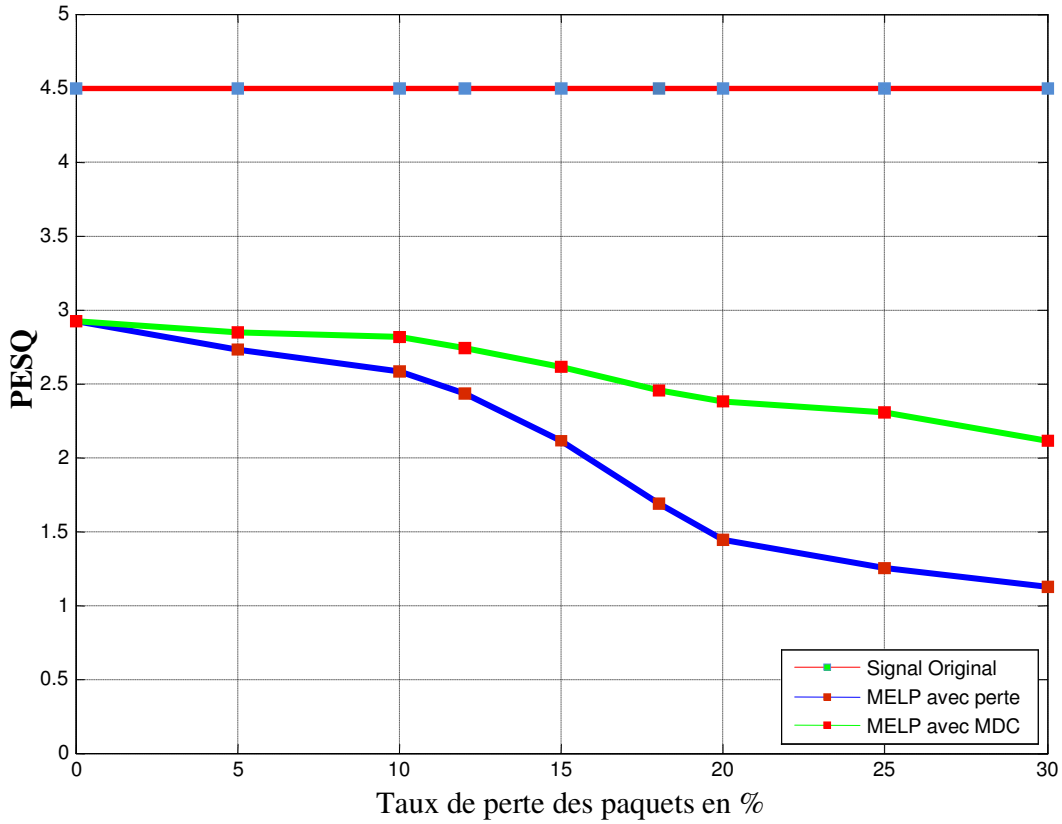


Fig. V.7. Evolution objective de la qualité en fonction des pertes par paquets dans le cas combiné des locuteurs et des locutrices.

V. 9. Interprétations des résultats

Nous avons simulé les pertes par dix taux, variant de 0% à 30%, ceci permet de bien voir l'évolution de la qualité et de la distorsion en fonction du taux de perte. D'après les figures (Fig. V.5), (Fig. V.6) et (Fig. V.7) nous remarquons que la qualité de la parole restituée avec la technique MDC, pour récupérer les trames perdues proposée est nettement améliorée par rapport à la parole reconstituée sans la MDC. Cette amélioration croît proportionnellement avec le taux de perte, pour un taux de perte de 30 %, ceci permet de bien voir l'évolution de la qualité (PESQ) en fonction du taux de perte.

Pour un taux de perte allant de 0 jusqu'à 5%, l'influence des pertes sur le score PESQ est moins observable, la qualité reste bonne. Pour un taux de perte supérieur, jusqu'à 15%, la qualité se dégrade continuellement jusqu'à ce que l'intelligibilité soit perdue et surtout pour les locutrices féminines.

D'après les tableaux (Tableau V.3), (Tableau V.4) et (Tableau V.5), nous observons la différence des scores PESQ, après l'utilisation de la MDC, le rehaussement de la qualité en moyenne est d'environ de 0.49 pour les locuteurs masculins, 0.60 pour le cas combiné et 0.58 pour les locutrices féminines.

- **Exemples de phrases**

Nous donnons des exemples de résultats où l'on voit le signal reconstitué après une perte de paquets. On observe notamment la correction apportée par l'utilisation de la MDC.

➢ Cas du locuteur masculin

La figure V.8 montre un résultat obtenu sur la phrase Ph1 intitulée ‘نمنم ماء اليوم‘, prise dans la base de données PAPE.

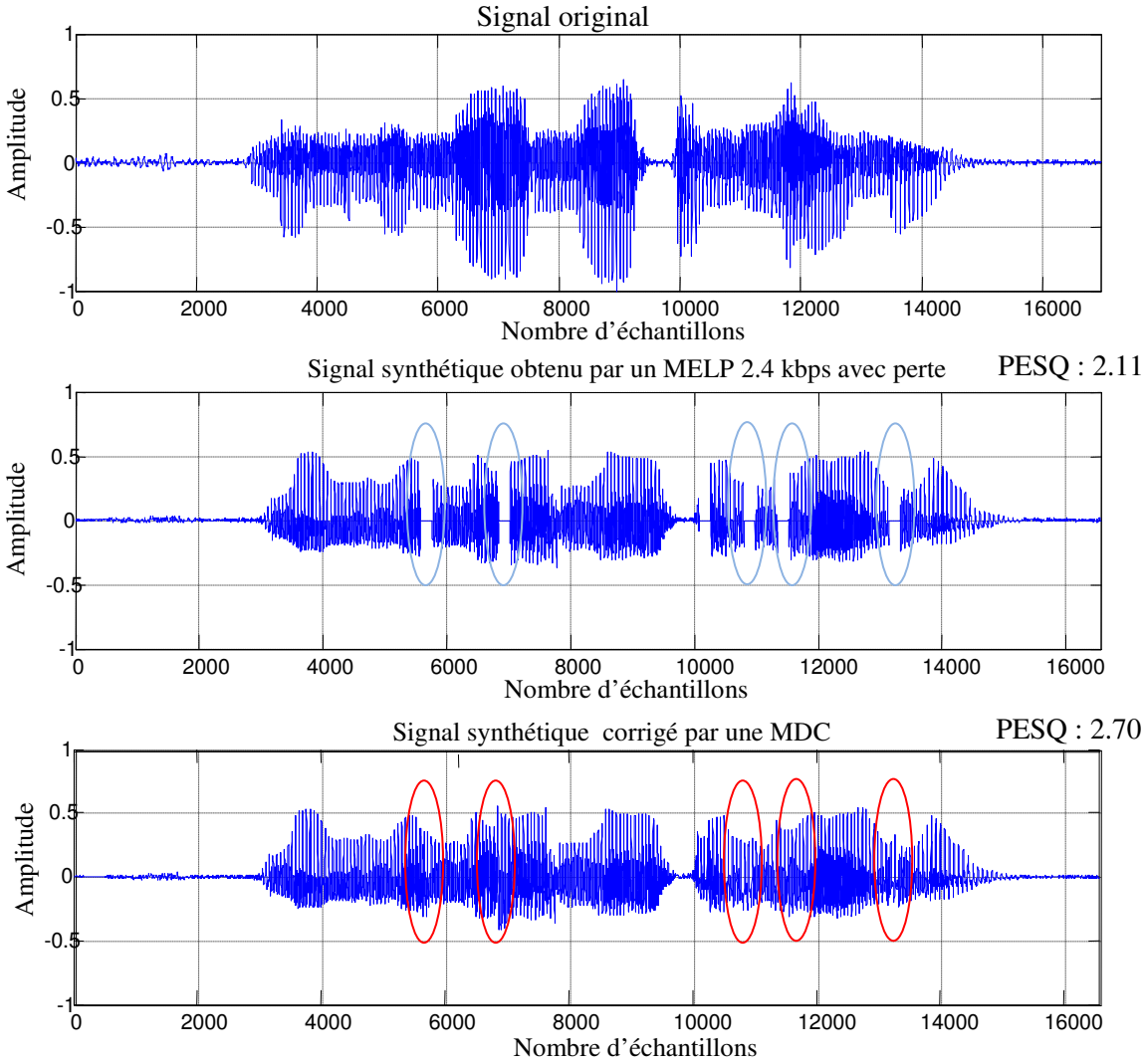


Fig. V.8. Résultats obtenus sur phrase ‘‘نمنم ماء اليوم’’. On observe :

a) signal original,

b) signal synthétique avec pertes de trames et

c) signal synthétique après correction avec une MDC.

- Cas du locuteur féminin

La figure V.9 montre un exemple de résultat obtenu sur la phrase Ph2 intitulée ولم يشفق عنها ‘, de la même la base de données PAPE prononcées par une locutrice féminine.

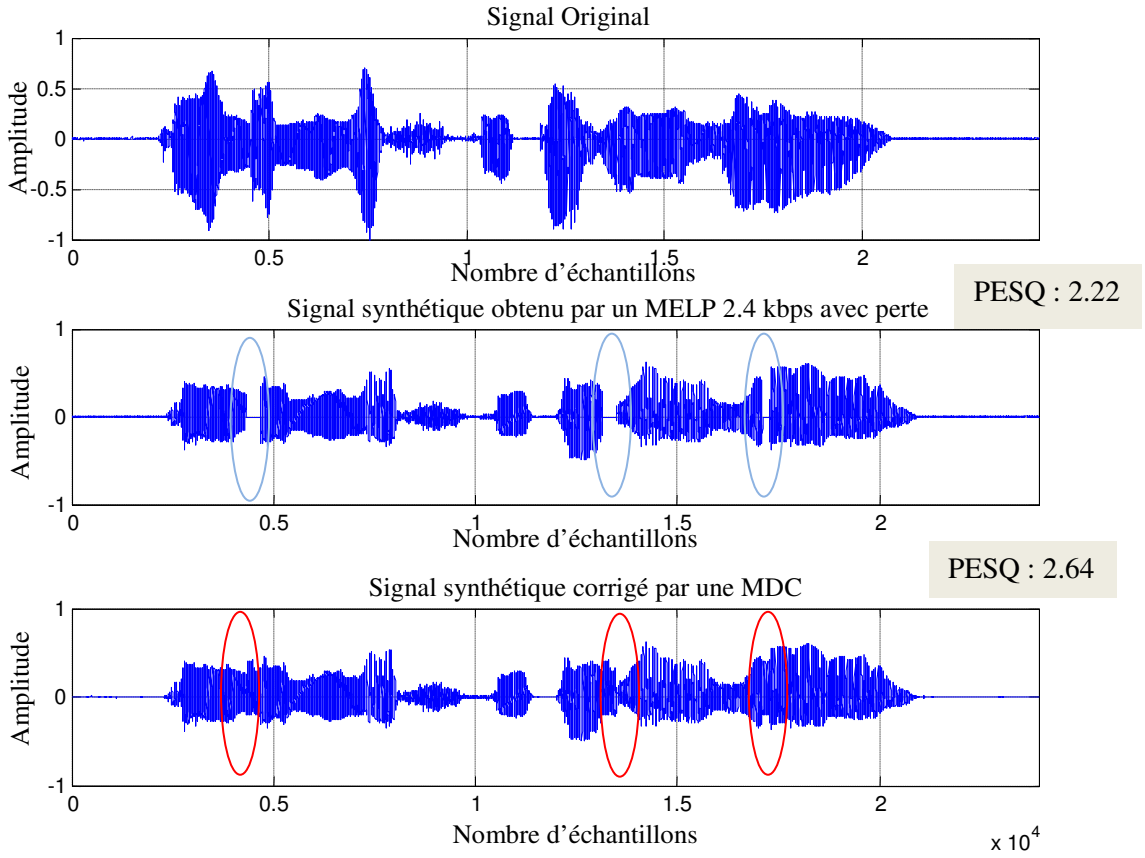


Fig. V.9. Résultats obtenus sur phrase ‘‘ولم يشفق عنها On observe :

a) signal original,

b) signal synthétique avec pertes de trames et

c) signal synthétique après correction avec une MDC.

Les corrections apportées procurent de bons résultats, c'est-à-dire qu'il ya un bon recouvrement des trames perdues, ces résultats graphiques sont certes significatifs, mais ils ne constituent pas un critère d'évaluation. Nous constatons que le meilleur jugement reste la mesure perceptuelle ou l'écoute.

Nous présentons ici des exemples d'agrandissement effectués autour des régions présentant des pertes de trames. Il s'agit des cas suivants : une seule trame perdue, deux trames successives perdues et trois trames successives perdues pour trois signaux différents.

Premier cas : Un agrandissement dans le cas d'une perte d'un seul paquet dans une région.

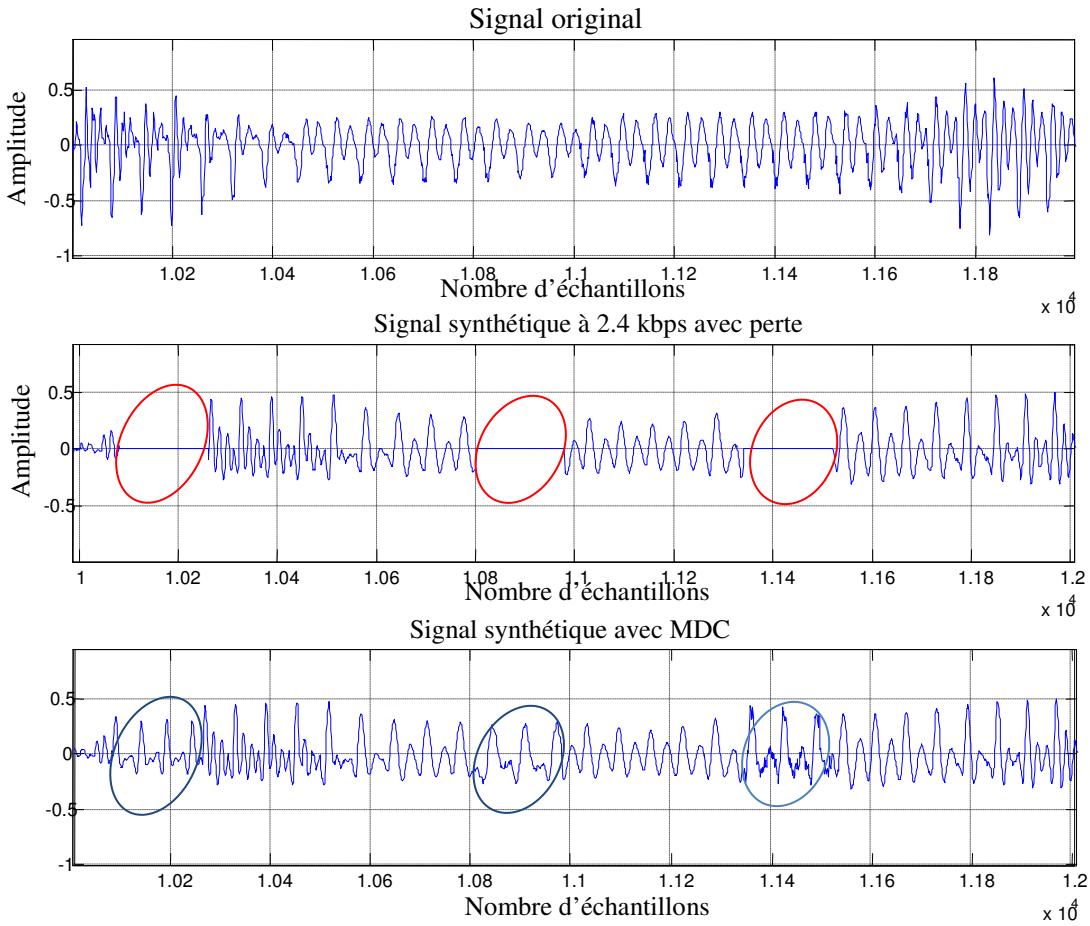


Fig. V.10. Exemple de portions du signal présentant des pertes de paquets :

a) signal original.

b) signal synthétique avec perte de trames.

c) signal synthétique après correction avec MDC.

Deuxième cas: Agrandissement effectué autour des régions présentant des pertes de deux trames consécutives.

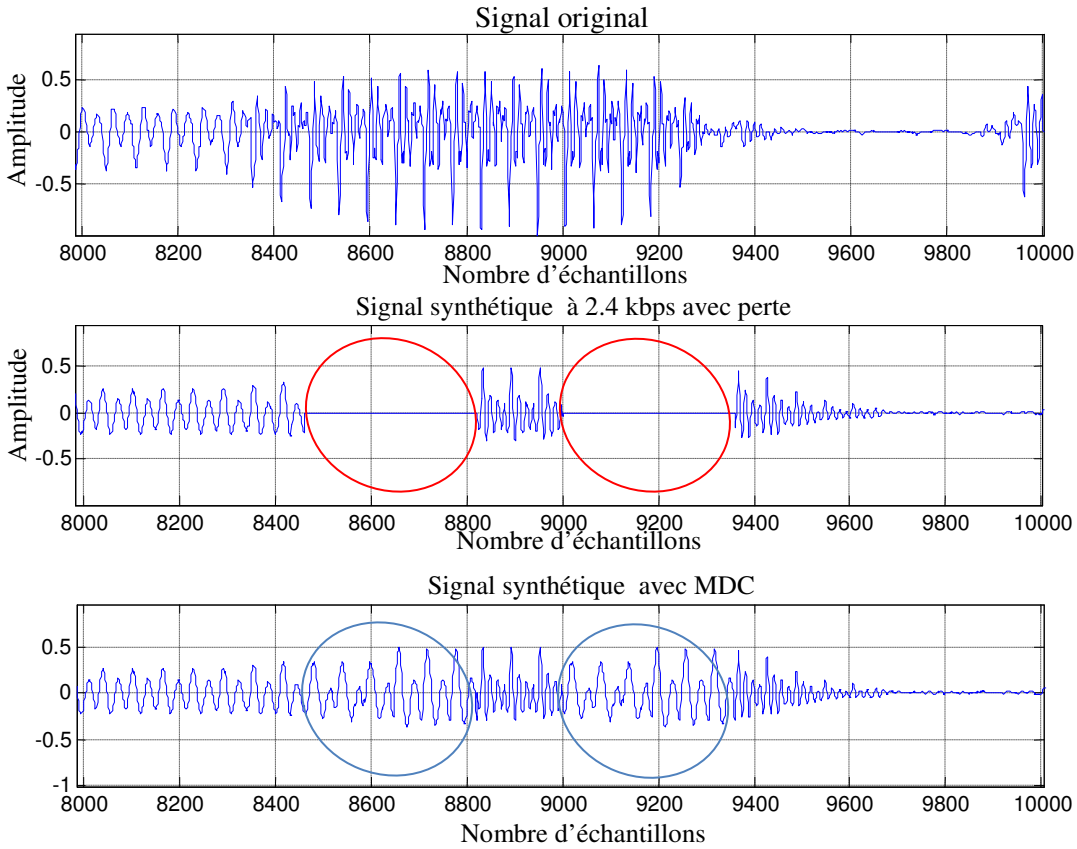


Fig. V.11. Exemple de portions du signal présentant des pertes de paquets

a) signal original.

b) signal synthétique avec perte de deux trames successives.

c) signal synthétique après correction avec MDC.

Troisième cas : Agrandissement effectué autour des régions présentant des pertes de trois trames successives.

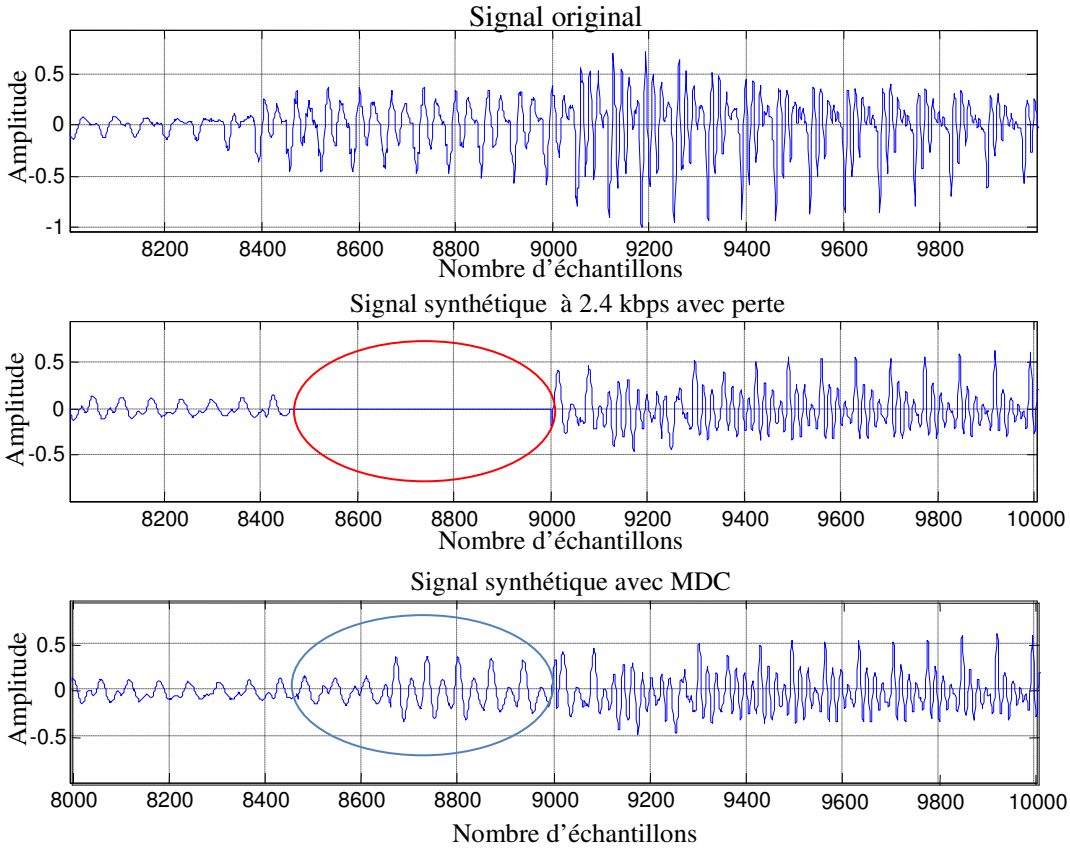


Fig. V.12. Exemple de portions du signal présentant des pertes de paquets :

a) signal original.

b) signal synthétique avec perte de trois trames successives.

c) signal synthétique après correction avec MDC.

Dans les figures V.8, figure V.9 et V.10, nous constatons le recouvrement des trames perdues pour différents cas de pertes de paquets. Pour le cas des zones voisées et après la correction, on voit bien que le voisement est conservé.

Conclusion

Dans ce chapitre nous avons vu l'évaluation des résultats obtenus par notre simulation sous MATLAB de la VoIP en utilisant deux codeurs MELP. Le fonctionnement correct démontré déjà par les tests d'évaluation effectués sur les différentes pertes. Le premier servant à la transmission sur un réseau IP de la parole codée sur 2.4 kbps. Le second fonctionnant à 1.2 kbps est ajouté au premier dans le même paquet pour servir au recouvrement des paquets perdus à l'aide de la technique MDC.

Nous pouvons conclure que la performance de la technique de codage par description multiple, apporte une amélioration importante de la qualité de la parole perceptuelle, spécialement lorsque le taux de perte est assez grand.

Conclusion générale

Dans ce travail, nous avons présenté une méthode originale utilisant deux codeurs harmoniques MELP dans la MDC, le premier servant à la transmission sur un réseau IP de la parole codée sur 2.4 kbps. Le second fonctionnant à 1.2 kbps est ajouté au premier dans un même paquet. Un paquet contient donc deux codeurs MELP organisés selon le principe d'une technique dite description multiple ou MDC, pour combattre les pertes de paquets lors de la transmission de la parole sur ce réseau IP.

Les résultats de notre simulation ont montré que notre méthode basée sur cette technique de dissimulation de perte pour une application VoIP rehausse la qualité de 0.68, pour des taux de perte allant jusqu'à 30%.

Ces résultats montrent que la MDC est efficace, et apporte une amélioration importante de la qualité de la parole, spécialement lorsque le taux de perte augmente. L'application de cette technique devient ainsi recommandée dans les réseaux perturbés et instables.

Nous avons montré que la redondance correctement ajoutée peut garantir une meilleure qualité de parole pour tout type de perte de paquets.

Cependant, et étant donné que toute réduction de débit d'un codeur de parole induit une augmentation de la charge de calcul, il est nécessaire de bien l'appréhender dans l'optique d'une application en temps réel.

Dans cette étude, nous avons obtenu un codeur qui fonctionne à 135 bits / trame de 22.5 ms correspondant à un débit de 6 kbps. On préconise cette solution pour remplacer avantageusement, le codeur CELP de la norme G.729 actuellement utilisé dans les applications de VoIP et qui présente un débit de 8 kbps sans MDC. En présence d'une MDC, ce débit devient 16 kbps.

On peut dire que ce travail nous a initiés à un domaine de recherche vaste qui est celui des télécommunications, associé aux applications multimédia. Nous avons maîtrisé beaucoup de techniques de codage et de traitement de la parole actuelle. Nous avons particulièrement travaillé sur la transmission de la voix à travers le réseau IP. Nous avons ainsi étudié les systèmes VoIP, leurs architectures, les protocoles utilisés et les problèmes rencontrés, notamment la perte de paquets. Nous avons contribué à cet axe en essayant de présenter une méthode originale pour remédier à ce problème de perte.

En perspective à ce travail, nous préconisons de comparer notre méthode à celles déjà existantes telles que la méthode d'entrelacement. En outre, il serait intéressant d'embarquer cette technique sur des processeurs de signal et d'effectuer la transmission en temps réel.

Bibliographie

[1] N. Moreau, “Techniques de compression des signaux”. Edition collection technique et scientifique des télécommunications, Paris-Tech, 1995.

[2] G. Baudoin, “Codage de la parole à bas et très bas débit transformation de la voix’’. Mémoire d’habilitation à diriger des recherches, université Marne La vallée, 2000.

[3] T**.** Dutoit, “Introduction au traitement automatique de la parole”. Faculté Polytechnique de Mons, TCTS Lab, 2000.

[4] C. E. Shannon, “A Mathematical Theory of Communication”. Bell System Technical Journal, 1948.

[5] J-P. Adoul et R. Lefebvre, “Speech coding and synthesis, chapter Wideband speech coding”. Elsevier, 1995.

[6] G. FUCHS, “Codage Audio Hiérarchique à Faibles Débits’’. Thèse de doctorat, université de Sherbrooke, Canada, 2007.

[7] G. Roy et B. Eng. “Low-rate analysis-by-synthesis wideband speech coding’’. Department of Electrical Engineering McGill université de Montreal, Canada, 1990.

[8] M. Jelinek, “Modélisation Spectrale et Compression de Parole à Bas Débit’’. Thèse de doctorat, université de Sherbrooke, Canada, 1998.

[9] N. Moreau, “Outils pour la compression application à la compression des signaux audio”. Edition collection technique et scientifique des télécommunications, Paris-Tech, 2007.

[10] S. Grassi “Optimized Implementation of Speech Processing Algorithms’’. Thèse de doctorat, université de Neuchatel, Switzerland, 1998.

[11] J. Makhoul, “Linear Prediction : A Tutorial Review’’. Proc. IEEE, pp. 561-580, 1975.

[12] B. Boudraa, “Analyse et Synthèse Multi-Impulsionnelle’’. Thèse de doctorat d’état, USTHB, 2006.

[13] B. Atal et J. Remde, “A new model for LPC excitation for producing natural sounding speech at low bit rates”. In Proc. ICASSP-82, pp. 616617, 1982.

[14] B.S. Atal and M.R. Schroeder, “Code-Excited Linear Prediction (CELP): high-quality speech at very low bit rates,” Proc. IEEE Int. Conf. on Acoustics Speech and Signal, Processing (ICASSP), 1985.

[15] A. Kondoz, “Digital Speech Coding for Low Bit Rate Communication Systems, Multi-Band Excitation Speech Coder’’. Wiley & Sons, 1994.

[16] Y.M Chung et S.U .Lee, “A comparison of two speech codes for digital mobile radio applications”. Speech Communication, vol. 11, N .1, pp 51-69, 1992.

[17] S. Ahmadi et A. S. Spanias, ‘‘Low-bit-rate speech coding based on an improved sinusoidal model’’. Speech Communication , 2001.

[18] A. Neubauer, J. Freudenberg et V. Kuhn, “Coding Theory: Algorithms, Architectures, and Applications ’’. Edition Wiley & Sons Ltd, 2007.

[19] A. McCree, K. Truong, E. B. George, T. P. Barnwell, et V. Viswanathan, “A 2.4 kbps MELP Coder Candidate for the New U.S. Federal Standard”. Proceedings of IEEE ICASSP, pp. 200-203, 1996.

[20] ITU-T, Rec. P.800 Methods for Subjective Determination of Transmission Quality, 1996.

[21] S .Wang, A. Sekey et A. Gersho, “An objective measure for predicting subjective quality of speech coders”. IEEE Journal Sel .Areas in communication, Vol 10, pp 819-829, 1992.

[22] W. Yang, M. Benbouchta et R. Yantorno, “Performance of the modified bark spectral distortion as an objective speech quality measure’’. ICASSP, vol. 1, pp. 541-544, 1998.

[23] W. Yang, M. Dixon et R. Yantorno, “A modified bark spectral distortion measure which uses noise masking threshold”, IEEE Speech Coding Workshop, pp. 55-56, 1997.

[24] A. Rix et P. Gray. “Non-intrusive Speech Quality Assessment”. UIT-T COM 12-D.48, 2001.

[25] M. Guéguin, G. Faucon, et V. Barriac. “Towards An Objective Model of the Conversational Speech Quality”. In Proc. ICASSP’06, 2006.

[26] J. Ellis, C. Pursell et J. Rahman, “The convergence of Voice, Video & Network Data” edition ACADIMIC PRESS, 2004.

[27] L. Ouakil et G. Pujolle, “ Téléphonie sur IP ’’. Edition groupe Eyrolles, 2008.

[28] A. Nagle, “Enrichissement de la Conférence audio en Voix sur IP au travers de l'amélioration de la qualité et de la spatialisation sonore’’. Thèse de doctorat, Paris-Tech, 2008.

[29] O. Hersent, D. Gurle, “L’essentiel de la VoIP’’. Dunod, 2005.

[30] G. Madre, “Application de la Transformée en Nombres Entiers à l’étude et au Développement d’un Codeur de Parole pour Transmission sur Réseaux IP’’. Thèse de doctorat Université de Bretagne Occidentale, 2004.

[31] S. Pracht et D. Hardman, “Voice quality in converging telephony and IP networks”. www.ednmad.com, 2000.

[32] G. Pujolle, “Les Réseaux”. Editions Eyrolles, 2008.

[33] J. Davidson, J. Peters, M. Bhatia, S. Kalidindi et S. Mukherjee “Voice over IP Fundamentals”. Edition Cisco Press, 2006.

[34] A. Raak, “Speech Quality of VoIP Assessment and Prediction’’. Edition Wily & Sons, Ltd, 2006.

[35] Thi Mai Trang Nguyen, “Service level negotiation for heterogeneous IP-based network’’. Thèse de doctorat, université de Paris VI, 2003.

[36] ITU-T, REC. “G.114 Temps de transmission dans un sens’’, 2003.

[37] O. Hersent, D. Gurle, J. Petit, “La Voix sur IP : Codecs, H.323, SIP, MGCP, déploiement et dimensionnement’’. Dunod, 2004.

[38] G. Frédéric, “Cours de Réseaux et systemes’’. Edition Cnam 2000.

[39] S. Krawczyk, “La téléphonie sur IP : Convergence, Enjeux et Composants”. Analyse IDC pour Cisco Systems, 2002.

[40] J.K. Wolf, A. Wyner et J. Ziv, “Source coding for multiple descriptions”. Bell Syst. Tech, vol. 59, No 8, pp. 1417–1426, 1980.

[41] T. Guionnet, “Codage robuste par descriptions multiples pour transmission sans fil d'information multimédia”. Thèse doctorat, université de Rennes1, 2003

[42] A. R. Reibman, H. Jafarkhani, Y. Wang, M. T. Orchard, et R. Puri, “Multiple description video coding using motion-compensated temporal prediction”. IEEE Trans. on Circ. and Syst. for Video Technology, vol. 12, pp. 193–204, 2002.

[43] J. Wang, X. Wu, S. Yu, et J. Sun, “Multiple descriptions with side informations also known at the encoder”. In Proc. IEEE International Symposium on Information Theory ISIT, pp. 1771–1775, 24–29, 2007.

[44] R. Puri, S. S. Pradhan, et K. Ramchandran, “N channel multiple descriptions: theory and constructions”. In Proc. Data Compression Conference DCC 2002, pp. 262–271, 2002.

[45] A. Gamal et T. Cover, “Achievable rates for multiple descriptions”. IEEE Trans. On Information Theory, vol. IT-28, pp. 851-857, 1982.

[46] Y. Wang, A. Reibman, T. Orchard et H. Jafarkhani, “An improvement to multiple description transform coding”. IEEE Transactions Signal Processing, vol. 50, no. 11, pp. 2843–2854, 2002.

[47] G. Kubin et W.B. Kleijn, “Multiple-description coding (MDC) of speech with an invertible auditory model”. In Speech Coding Proceedings, 1999 IEEE Workshop on, pp. 81-83, 1999.

[48] H. Coward, R. Knopp, et S. D. Servetto, “On the performance of multiple description codes over bit error channels”. Proc. IEEE International Symposium on Information Theory, pp. 240, 24–29, 2001.

[49] A. Mohr, E. Riskin et R. Ladner, “Generalized multiple description coding through unequal loss protection”. Proc. IEEE Intl. Conf. on Image Processing, ICIP'99, 1999.

[50] M. Rui et F. Labeau, “Error-Resilient Multiple Description Coding’’. Proceedings of the IEEE, vol 56, No 8, 2008.

[51] V. K. Goyal, “Multiple description coding: Compression meets the network’’. IEEE Signal Processing Magazine, vol. 18, pp. 74.93, 2001.

[52] V. K. Goyal, J. Kovaeevi, et J. A. Kelner, “Quantized frame expansions with erasures”. Applied and Computational Harmonic Analysis, vol. 10, pp. 203–233, 2001.

[53] X. Zhong et B.H. Juang, “Multiple description speech coding with diversities”. In ICASSP, vol. 1, pp. 177–180, 2002.

[54] O. Crave, “Approches théoriques en codage vidéo robuste multi-terminal”. Thèse doctorat, université de Rennes, 2008.

[55] W. C. Chu, “Speech Coding Algorithms Foundation and Evolution of Standardized Coders’’. Wiley & SONS INC, 2004.

[56] A. McCree, K. Truong, E. B. George, T. P. Barnwell, et V. Viswanathan, “A 2.4 kbps MELP Coder Candidate for the New U.S. Federal Standard”. Proceedings of IEEE ICASSP, pp. 200-203, 1996.

[57] Y. Medan, E. Yair, et D. Chazan, “Super Resolution Pitch Determination of Speech Signals”. IEEE Transactions on Signal Processing, Vol. 39, No. 1, 1991.

[58] V. Krishnan, “A Framework for Low bit-rate Speech Coding in Noisy Environment’’. Thèse de doctorat, Institute de Technologies de Georgia, 2005.

[59] A. McCreel, K.Truong, E. Bryan George, T. P. Barnwell et V. Vzswanathanl, ‘‘A 2.4 kbps MELP coder candidate for the new U. S. Federal Standard’’, Transactions on Speech and Audio Processing, 1996.

[60] A. McCree, et T. Barnwell-III, ‘‘A mixed excitation LPC vocoder model for low bit rate speech coding’’. IEEE Transactions on Speech and Audio Processing, vol. 3, pp. 242-250, 1995.

[61] P. Kabal et R. Ramachandran, “The Computation of Line Spectral Frequencies Using Chebychev Polynomials”. IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-34, No. 6, pp. 1419-1426, 1986.

[62] T. Wang, K. Koishida, V. Cuperman et A. Gersho, “A 1200 bps Speech Coder Based on MELP’’. Proceedings of IEEE, pp.III-1375-1378, 2000.

[63] L. Arslan, A. McCree, et V. Viswanathan, “New Methods for Adaptive Noise Suppression ’’. Proceedings of IEEE ICASSP, pp. 812-815, 1995.

[64] L.M. Supplee, R. P. Cohn, J. S. Collura, A. McCree, ‘MELP: The new federal standard at 2400 bps”. IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP, pp.II-1591-1594, 1997.

[65] Edward J. Daniel et Keith A. Teague, ‘‘Federal Standard 2.4 kbps MELP over IP’’. IEEE Transactions of Acoustics, Speech and Signal Processing, 2000.

[66] T.T.Teo, E.C.Tan, ‘‘Real time implementation of MELP vocoder’’. Journal of The Institution of Engineers, Singapore Vol. 44 Issue 3, 2004.

[67] ITU-T, Recommendation P.862, “Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs”. 2001.

[68] R. Beuran, “ Mesure de la qualité dans les réseaux informatiques’’. Thèse de doctorat, université de Jean Monnet St Etienne, 2004.

[69] M. Boudraa, B. Boudraa, B. Guérin. “Twenty lists of Ten Arabic Sentences for Assessment’’. Acustica, vol. 86, pp. 870-882, 2000.

[70] E. Orozco, S. Villette et A. Kondoz, “Multiple Description Coding for Voice over IP using Sinusoidal Speech Coding’’. Proceedings of IEEE, ICASSP, 2006.